Volume 21, Part 1 *May 1968*

THE BRITISH JOURNAL OF MATHEMATICAL & STATISTICAL PSYCHOLOGY

EDITED BY R. J. AUDLEY

WITH THE ASSISTANCE OF CYRIL BURT

TAYLOR & FRANCIS LTD
RED LION COURT, FLEET STREET, LONDON, E.C.4

40*s. net* $7.45

Book Review Editor: A. R. JONCKHEERE

Abstracts Editor (Computer Programs): H. J. HALLWORTH

Editorial Board:

C. Banks
M. S. Bartlett
G. H. Bower
D. E. Broadbent
J. Brown
V. R. Cane
L. J. Cronbach
C. I. Howarth
A. R. Jonckheere
D. N. Lawley
R. D. Luce
S. Messick
E. A. Peel
L. S. Penrose
J. Fraser Roberts
W. Stephenson
A. Stuart
P. E. Vernon

The British Journal of Mathematical and Statistical Psychology is issued by the British Psychological Society. Details of subscriptions to this and other Journals of the Society are given inside the back cover.

Papers for publication and books for review should be sent to the Editor, Professor R. J. Audley, Department of Psychology, University College London, Gower Street, W.C.1.

Abstracts of computer programs should be sent to Mr. H. J. Hallworth, Department of Education, University of Birmingham, Birmingham 15. Articles on computers, computer languages and programming will also occasionally be published and should be addressed to Mr. Hallworth.

PREPARATION OF PAPERS

Contributors are asked to send their papers in a form which is ready for submission to the printer: that is to say, the arrangement of the material should follow that observed in a current issue of this *Journal*. Each article should be headed with a brief abstract. Each section should have a short cross-heading, in capitals; and the whole should conclude with a summary embodying the main conclusions reached. Special attention should be paid to such details as correct spelling, grammar, capitalization, numbering, etc. In preparing manuscripts and in correcting proofs authors should conform, so far as possible, with the 'Recommendations to Authors' set out in *The Printing of Mathematics* by T. W. Chaundy, P. R. Barrett, and Charles Batey (Oxford University Press, 1954, ch. 2, pp. 21–73), or to those given in the paper on 'Setting Mathematics' (Arthur Phillips, *Monotype Recorder*, 60, iv, 1956).

Owing to the present conditions for printing, the cost of setting up tables and algebraic formulae which require hand composition is very much greater than the cost of machine composition. Contributors are therefore advised to arrange their material so that it can be printed as economically as possible. Often numerical results and algebraic formulae can, with a little simplification, be readily adapted for purposes of machine composition. Authors must ensure that formulae are clear, particularly when these contain subscripts.

Tables, photographs and figures which are costly to print should be used only when they save considerable space or are essential for effective presentation; and in no case should both a figure and table be used to convey the same information. Each table or figure should be numbered and on a separate sheet with its approximate position indicated in the typescript.

Drawings should be in Indian ink on heavy unruled paper. Lettering on graphs, etc. is best indicated in a tracing placed over the diagram. Diagrams are normally reduced in reproduction and allowance must be made for a corresponding reduction in the thickness of lines and the spaces between them. The numbered legend of a figure should be typewritten on a separate sheet of paper, not lettered on the drawing.

The object of the caption for a table or the legend for a figure should be to make the table or the figure comprehensible without reference to the text.

References in the text should be indicated by dates in parentheses after the authors' names and correspond to an alphabetical list at the end of the paper. The list must follow the layout illustrated in the three following examples:

ALLPORT, G. W. & VERNON, P. E. (1931). A test for personal values. *J. abnorm. soc. Psychol.* **26,** 231–248.

GALTON, F. (1883). *Inquiries into Human Faculty and its Development.* London: Macmillan.

RÉVÉSZ, G. (1951). Le problème du génie. *Année psychol.* **50,** 83–96.

World list abbreviations are used for journals. The Oxford Dictionary is taken as the authority on spelling. For other general guidance, authors are advised to consult *Suggestions to Authors* issued by the Standing Committee on Publications of the British Psychological Society and obtainable, price 5*s.* (U.S.A. $1.00) post free, from Cambridge University Press.

In future authors will receive one complete slip-proof of their papers, but no page proof. They will be expected to pay the cost of all corrections for which they themselves are responsible.

All manuscripts should be submitted in duplicate.

PROGRAM ABSTRACTS

An abstract of a computer program, normally of 300 to 500 words in length, should state the sequence of standard mathematical or statistical operations that are performed, the language in which the program is written, the approximate running time and should either give an address from which the program can be obtained (in tape or card form) or state where data may be sent for processing. Authors are referred to a current issue of the *Journal* for examples of an acceptable abstract. In most cases, the abstract should be accompanied by an actual program in the form of a deck of cards or a paper tape together with the appropriate listing information, also any other instructions which are necessary for the preparation of data or the running of the program.

Vol. 21
Part 1
1–19

The British Journal of Mathematical and
Statistical Psychology

May
1968

32

GOVT. OF W
LIBR
CALCUT

• MULTIPLE REINFORCEMENT EFFECTS IN SHORT-TERM MEMORY

By J. W. Brelsford, Jr.,[1] R. M. Shiffrin and R. C. Atkinson
Stanford University, Stanford, California

A continuous memorizing situation was studied in which test and study trials alternated throughout an experimental session. The items studied were paired-associates. The interval between study and test for a particular item was randomly determined; and an item was given one, two, three or four reinforcements. A quantitative model is proposed which has two memory stores: a short-term store in which the subject generates a carefully controlled rehearsal scheme of fixed length, and a long-term store in which information is accumulated and lost. A large number of theoretical predictions of the model were verified quantitatively by the data, which confirm results of previous experiments, and support the hypothesis that highly structured rehearsal schemes play a major role in many short-term memory and learning situations.

1. Introduction

Atkinson & Shiffrin (1968) have proposed a multi-process model of memory that has been employed successfully to describe the effects of a variety of experimental variables. These include: list length, presentation rate and confidence ratings in experiments involving a discrete, independent trial procedure; and mode of rehearsal, size of stimulus set, and judgements of recency in continuous memory experiments. In all of these experiments, items to be recalled were presented for a single period of study with a subsequent test period. In so far as a given item never received more than a single study period (or reinforcement), these experiments may be said to involve memory rather than learning. The present study extends the earlier work into the area of paired-associate learning by giving items varying numbers of reinforcements. An attempt will be made in this paper to explain the results with a model incorporating only a minor extension of the theory used in the previous work.

The task employed in the experiment involves a continuous presentation technique which makes it possible to study the learning process under conditions that are quite uniform and stable throughout the course of an experiment. (This technique is very similar to one employed by Yntema & Mueser (1960, 1962).) In essence the task requires the subject to keep track of the randomly changing responses to eight different stimuli. The eight stimuli are chosen at the start of a session and used throughout that session. A session is begun by presenting for study each of the eight stimuli with associated responses.

---

[1] Now at Yale University

Bureau of Ednl. & Psyl. Research S.E.
(S. C. E. R. T.)
Date ....
Acc .. 896

Following this initial study phase, there is a continuous series of trials, each trial consisting of a test phase followed by a study phase. During the test phase a stimulus is randomly selected from among the set of eight stimuli, and the subject tries to recall the response *last*, or *most recently*, associated with that stimulus. Following the test (and the subject's attempted recall) the study phase of the trial occurs. In the study phase, the same stimulus presented in the preceding test phase is presented, sometimes re-paired with the response that was previously correct, or sometimes paired with a new response; in any case, the subject must study and try to remember the presented pair. In order to distinguish between a particular stimulus and a particular stimulus–response pair, a convention is henceforth adopted that an ' item ' will refer to a particular stimulus–response pair. The number of reinforcements for a given item is determined probabilistically but never exceeds four.

The number of trials intervening between the study and test of a given item will be referred to as the ' lag ' for that item. Thus, if the test occurs immediately following the study period, the lag is zero. If one trial intervenes (involving test and study on another stimulus), then the lag is one; and so on. Since the stimulus tested is chosen randomly from the set of eight stimuli on each trial, the lag between study and the next test is distributed geometrically with a parameter of 1/8. The task of the subject is simply to remember the current responses assigned to the eight different stimuli. Learning is involved because at the time of testing some of the stimulus–response pairs have had multiple reinforcements, i.e. any given study period may involve the first, second, third or fourth reinforcement of a particular stimulus–response pair. Therefore, the subsequent test of that pair will involve a test after one, two, three or four reinforcements, respectively. The primary dependent variable is the probability of a correct response as a function of lag and the prior number of reinforcements.

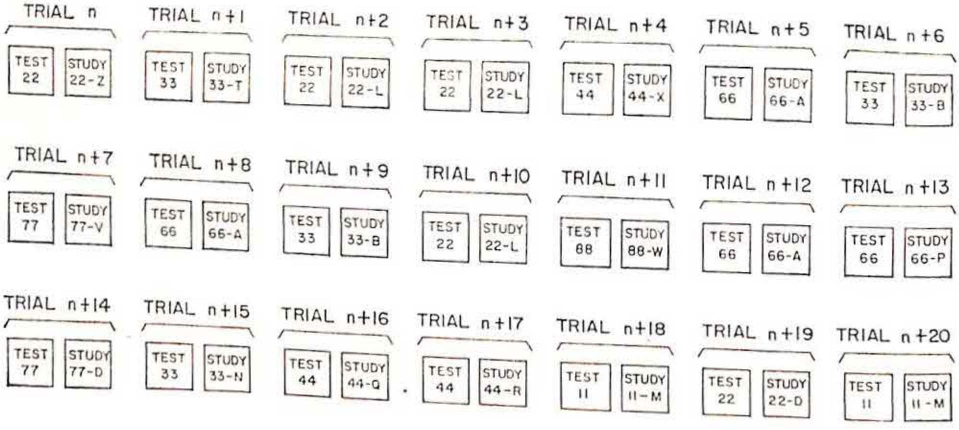


FIGURE 1. A sample sequence of trials.

Before proceeding to the model it may be helpful to illustrate the experimental procedure in some detail. Fig. 1 presents a sample sequence of trials from trial n to trial $n+20$. The stimuli were selected from the set of two digit numbers, and the responses were selected randomly from the alphabet. Various events of the type to be considered later are illustrated in Fig. 1. On trial $n+2$ stimulus 22 is paired with a new response, L, and assigned three reinforcements, the first occurring on trial $n+2$. The second reinforcement occurs on trial $n+3$ after a lag of zero. After a lag of six, the third reinforcement occurs on trial $n+10$. After a lag of eight, stimulus 22 is re-paired with a new response on trial $n+19$. Stimulus 33 is sampled for test on trial $n+6$ and during the study phase is assigned the new response, B, which is to receive two reinforcements, the second on trial $n+9$. Stimulus 44 is tested on trial $n+4$, assigned the new response, X, which is to receive only one reinforcement; thus, when 44 is presented again on trial $n+16$ it is assigned another response which by chance also is to receive only one reinforcement, for on the next trial 44 is studied with response Q. Note also what constitutes a correct response. For example, when stimulus 33 is presented for test on trial $n+6$, the correct response is T; when presented for test on trials $n+9$ and $n+15$, the correct response is B; and when presented again, the correct response will be N.

2. MODEL

The theory postulates three memory states: a very short-lived memory system called the sensory register; a temporary memory state called the short-term store (STS); and a more permanent memory storage state called the long-term store (LTS). A consideration of the time intervals and procedure used in the present study leads us to assume that every item is accurately recorded in the sensory register and then at once transferred to STS for active consideration by the subject. In what follows, then, the sensory register plays no part and attention will be restricted to STS and LTS.

Short-term Store

Information entering STS resides there for a short period, perhaps of the order of 15 sec., before being lost completely; this period may be extended indefinitely, however, by processes controlled by the subject, such as rehearsal, coding, and so forth. The memory in the model is imperfect, nevertheless, because only a very limited amount of information may be maintained in STS at any one time. An immediate test of any information that has entered STS (within a few seconds of study, say) will result in correct retrieval of that information. In the present study every item is assumed to enter STS and a test at lag zero occurs within several seconds of study; therefore, the probability of a correct response for any item tested at lag zero is unity. For lags of one or greater the time intervals in the present task are such that items will be lost from STS unless some sort of rehearsal mechanism is invoked to maintain them.

An analysis of subjects' reports and a consideration of similar experiments (Atkinson, Brelsford & Shiffrin, 1967) lead us to propose a very specific, quite orderly, rehearsal scheme called the ‘ rehearsal buffer ’ or ‘ buffer ’.

Rehearsal Buffer The buffer is a rehearsal scheme in which a fixed number, r, of homogeneous items are rehearsed at any one time during an experimental session. It is assumed that the series of study items at the start of each experimental session fills the buffer and that the buffer stays filled thereafter. The size of the buffer, r (defined as the number of items held simultaneously), depends upon the nature of the items and thus must be estimated. Once the buffer is filled, each new item that enters causes one of the items currently in the buffer to be lost (i.e. an item currently undergoing rehearsal is eliminated to make room for the new item). It is assumed that a correct response is always given if an item is in the buffer at the time it is tested.

Consideration will now be given to the decision rules by which the subject enters new items into the buffer, and thereby eliminates items currently in the buffer. In order to do this, it is necessary to distinguish theoretically between two kinds of items presented for study: *O*-items and *N*-items. An *O*-item (old-item) is an item presented for study whose stimulus is already a component of an item currently in the buffer. An *N*-item (new-item) is a presented item whose stimulus is not at that time in the buffer.

RULE 1. *Any O-item presented for study is automatically entered into the buffer and replaces there the item with the same stimulus as the presented item.*

There are several reasons for this rule. If the presented item is receiving its second, third or fourth reinforcement, then the item replaced is identical to the one presented and it does not really matter whether we speak of replacement or not, i.e. the state of the buffer remains unchanged in any case. If, on the other hand, the presented item is receiving its first reinforcement, then the item currently in the buffer with the same stimulus will have a different response member. In fact, this response will now be an incorrect response to that stimulus; the subject must therefore change it to the new, correct response, lest he start rehearsing incorrect information.

RULE 2. *When an N-item is presented for study, the subject bases his decisions upon the result of the immediately preceding test phase of the trial: if on the test phase the subject had correctly retrieved from long-term store the item being presented for study, then the item is not entered into the buffer. Otherwise, the item is entered into the buffer with probability* α*; if the item enters the buffer, then the item eliminated to make room is chosen randomly (i.e. each item currently in the buffer has probability* $1/r$ *of being the one eliminated).*

The reasons for this rule are again straightforward. If a correct retrieval of the presented item had just been made from LTS, then the subject has reason to believe he already ‘ knows ’ it, and will not attempt to rehearse it further (in this case, of course, the presented item is receiving its second, third or fourth reinforcement). Note that a correct LTS retrieval of the presented item (and not just a correctly-guessed response or a zero-lag response) is necessary for the subject to prohibit entry into the buffer. If the presented N-item is not retrieved from LTS (which is always the case if the item is being given its first reinforcement) then the subject does not ' know ' it, and rehearsal becomes desirable. On the other hand, there undoubtedly is a certain amount of effort involved in rearranging the buffer; in addition, the item which must be removed from the buffer to make room has not yet been tested. In consideration of these conflicting tendencies it seems appropriate to let the entry probability be a parameter, α, to be estimated. Finally, once an item is entered into the buffer, it is necessary to decide which item already there is to be removed. In previous studies using fixed lists, it has proved useful to postulate a tendency for the oldest items in the buffer to be the first eliminated. In all work using the continuous presentation technique of the present study, however, it has been most accurate to postulate a random choice of the item to be removed. Fairly direct evidence supporting this assumption will be presented later in conjunction with Fig. 2.

Long-term Store

LTS is viewed as a memory state in which information accumulates for each item. The term ' information ' is not used here in a technical sense. It refers to codes, mnemonics, images or anything else the subject might store that would be retrievable at the time of test. It is assumed that information about an item may enter LTS only during the period that an item resides in the buffer. The status of an item in the buffer is in no way affected by transfer of information to LTS. Whereas recall from the buffer was assumed to be perfect, recall from LTS is not necessarily perfect and usually will not be. At the time of test on an item, a subject gives the correct response if the item is in STS but if the item is not found in STS, the subject searches in LTS. This LTS search is called the *retrieval process*. Two features of the LTS retrieval process must be specified. First, it is assumed that the likelihood of retrieving the correct response for a given item improves as the amount of information stored concerning that item increases. Second, the retrieval of an item gets worse, the longer the information has been stored in LTS. This could result from autonomous decay or active interference from other information being stored in LTS.

It will be specifically assumed in this paper that information is transferred to LTS at a constant rate, θ, during the period that an item resides in the buffer; θ is the transfer rate per trial. Thus, if an item remains in the buffer for exactly j trials, then that item accumulated an amount of information in LTS equal to $j\theta$. Next, it is assumed that each trial following the trial on which an item is knocked out of the buffer causes the information stored in LTS for that item to decrease by a constant proportion τ. Thus, if an item receiving its first reinforcement were knocked out of the buffer at trial j, and i trials intervened between the original study and the test on that item, the amount of information stored in LTS at the time of test would be $(j\theta)\ (\tau^{i-j})$. In this experiment an item receiving two, three or four reinforcements may enter and leave the buffer two, three or four times. When the item is in the buffer the θ-process is activated, and when not in the buffer the τ-process takes over.

The probability of a correct retrieval of an item from LTS is now specified. If the amount of information in LTS at the moment of test is zero, then the probability of retrieving the correct response should be zero. As the amount of information increases, the probability of a retrieval should increase toward unity. σ_I is defined as the probability of a correct response for an item which is not in the buffer but has accumulated an amount of information in LTS equal to I at the time of test. Considering the above specifications on the retrieval process,

$$\sigma_I = 1-(1-g)\exp(-I), \tag{1}$$

where g is the guessing probability and, in the present experiment, is 1/26 since there were 26 response alternatives.

As already stated, items that have received two, three or four reinforcements may have been in and out of the buffer at various times and the expression for I becomes fairly complex. Suppose, for example, an item has been in the buffer i_1 trials, then out of the buffer j_1 trials, then in i_2, out j_2, in i_3, out j_3, and then tested. At that point

$$I = \{[(i_1\theta)\tau^{j_1} + i_2\theta]\tau^{j_2} + i_3\theta\}\tau^{j_3}.$$

As part of the calculation of the probability correct at lag k for a three reinforcement item, one would then have to consider all combinations of i's and j's, and compute the conditional probability of each one. It should be clear that this direct method for calculating predictions, although possible, is both formidable and inefficient. Instead, predictions from the model were generated using Monte Carlo methods; the procedure is described in detail later.

At this point it would be appropriate to review the sequence of events and decisions proposed in the model. A trial begins with the presentation of a stimulus for test. The subject first checks STS; if the test was zero lag then the item is present in STS and a correct response is made. Next, or at the same time, the buffer is checked; if the stimulus is in the buffer, then again a correct response is made. If the stimulus is not in the buffer or STS, then LTS is searched, and the probability of retrieving the response correctly is an exponential function of the amount of information then in LTS about the item. If all of these searches fail, then the subject guesses randomly. Next comes the study phase of the trial. The stimulus just tested is now presented again paired with a response for study. In order to determine whether to rehearse this item, the subject refers to the two decision rules given earlier (concerning O- and N-items). It should be noted that the model has four parameters to be estimated: r, the buffer size; α, the buffer entry probability; θ, the transfer parameter; and τ, the long-term decay parameter. The model is identical to that used in previous studies (cf. Atkinson, Brelsford & Shiffrin, 1967) with the sole addition that an item retrieved from LTS is never entered into the buffer, a state of affairs never arising in experiments where each item receives only one reinforcement.

3. EXPERIMENT

Method

Subjects The subjects were 20 students from Stanford University who received $2 per experimental session. Each subject participated in at least 10 sessions.

Apparatus The experiment was conducted in the Computer-Based Learning Laboratory at Stanford University. Control functions were performed by computer programs running in a modified PDP–1 computer manufactured by the Digital Equipment Corporation and under the control of a time-sharing system. The subject was seated at a cathode-ray tube display terminal. Stimuli were displayed on the face of the cathode-ray tube (CRT); responses were made on an electric typewriter keyboard located immediately below the lower edge of the CRT. (For a more detailed description of the apparatus see Atkinson, Brelsford & Shiffrin, 1967.)

Stimuli and responses The stimuli were eight two-digit numbers randomly selected for each subject and session from the set of all two-digit numbers between 00 and 99. Once a set of eight stimuli was selected for a given session, it was used throughout the session. Responses were letters of the alphabet, thus fixing the guessing probability of a correct response at 1/26.

Procedure Every session began with a series of eight study trials; one study trial for each stimulus used in the session. On these trials each of the eight stimuli was paired with a response selected randomly. There were no restrictions on repetition of responses. After the initial study trials the session involved a series of consecutive trials, each consisting of a test phase followed by a study phase. On each trial a stimulus was randomly selected for testing, and the same stimulus was then presented for study. During the study phase of the trial the stimulus was sometimes re-paired with a new response and sometimes left paired with the old response. To be precise, when a particular stimulus–response pair was presented for study the first time, a decision was made as to how many reinforcements (study periods) it would be given; it was given either one, two, three or four reinforcements with probabilities 0·30, 0·20, 0·40 and 0·10, respectively. When a particular stimulus–response pair had received its assigned number of reinforcements, its stimulus was then randomly re-paired with a new response on the next study trial, and this new item was assigned a specific number of reinforcements using the above probability distribution. Reference to Fig. 1 should clarify this procedure. The subject was instructed to respond on the test phase of each trial with the letter that was *last* studied with the stimulus being tested. Since the stimulus selected for testing was chosen randomly on each trial, the distribution of the lag from study to test was geometric with a parameter of 1/8.

The temporal arrangement for the eight initial study trials was such that each study trial lasted for 3 sec. with a 3 sec. inter-trial interval. Each trial of the session proper involved the following sequence of events: (1) The word *test* appeared on the upper face of the CRT. Beneath the word *test* a randomly selected member of the stimulus set appeared. This test portion of a trial lasted for 3 sec., during which subjects were told to respond with the last response that had been associated with that stimulus, guessing if necessary. (2) The CRT was blacked out for 2 sec. (3) The word *study* appeared on the upper face of the CRT for 3 sec. Below the word *study* a stimulus–response pair appeared. The stimulus was the same one used in the preceding test portion of the trial. Depending upon the reinforcement schedule, the response was either the one that had previously been correct or a new one. (4) There was a 3 sec. inter-trial interval before the next trial. Thus a complete trial (test plus study) took 11 sec. A subject was run for 220 such trials during each experimental session. For each subject and session the entire sequence of presentations, the subjects' responses, and response latencies were recorded.

4. Results

In order to eliminate warm-up effects from the data, the first session for each subject and also the first 25 trials of each subsequent session are not included in any of the analyses; otherwise, in what follows, the results are pooled over all sessions and subjects.

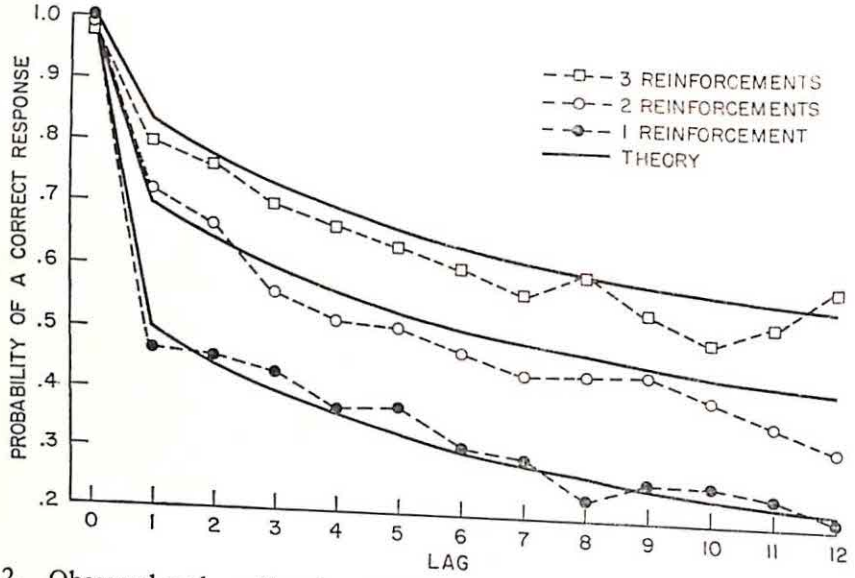
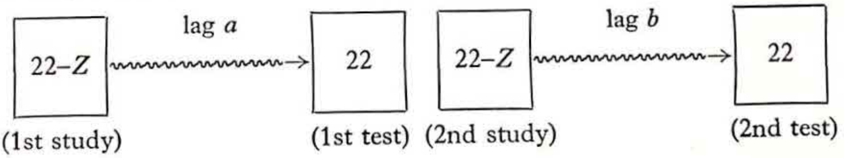


FIGURE 2. Observed and predicted probabilities of a correct response as a function of lag for items tested following their first, second or third reinforcement.

Fig. 2 presents the probability of a correct response as a function of lag for items tested after their first, second and third reinforcements. The number of observations is weighted toward the short lags, and also toward the smaller numbers of reinforcements. The short lags have more observations because the distribution of lags is geometric (with parameter 1/8). The smaller numbers of reinforcements are weighted because an item receiving k reinforcements provides data for every number of reinforcements less than or equal to k. For example, the one reinforcement lag curve contains not only data from items given just one reinforcement, but also data from the first reinforcement of items given two, three and four reinforcements. No data are presented for the four reinforcement conditions because of the small number of observations. What is graphed is the probability of a correct response to an item that received its jth reinforcement and was then tested after a lag of n trials. The data are presented for values of n ranging from 0 to 15 and for j equal to 1, 2 and 3.

The curves in Fig. 2 exhibit a consistent pattern. The probability correct decreases regularly with lag, starting at a higher value on lag 1 the greater the number of prior reinforcements. The forms of the 1-reinforcement curve and the probability values are quite similar to those found in a previous study with a similar procedure (Atkinson, Brelsford & Shiffrin, 1967) in which only one reinforcement was given to each item. There is one immediate inference that can be derived from these curves: accepting for the moment the rest of the model, the assumption that items to be lost from the buffer are chosen randomly is reasonably accurate. If there were a sizable tendency for the oldest items in the buffer to be lost first, then the curves would exhibit a pronounced S-shaped effect (Atkinson & Shiffrin, 1968).

For a deeper analysis of this experiment, certain dependencies masked by the curves of Fig. 2 need to be considered. For example, the probability of a correct response to an item that received its second reinforcement and was tested at some later trial will depend on the number of trials that intervened between the first and second reinforcements. To clarify this point consider the following diagram:



Item 22–Z is given its first reinforcement, tested at lag a and given a second reinforcement, and then given a second test at lag b. For a fixed lag b, the probability of a correct response on the second test will depend on lag a. In terms of the model it is easy to see why this is so. The probability correct for an item on the second test will depend upon the amount of information about it in LTS. If lag a is extremely short, then there will have been very little time

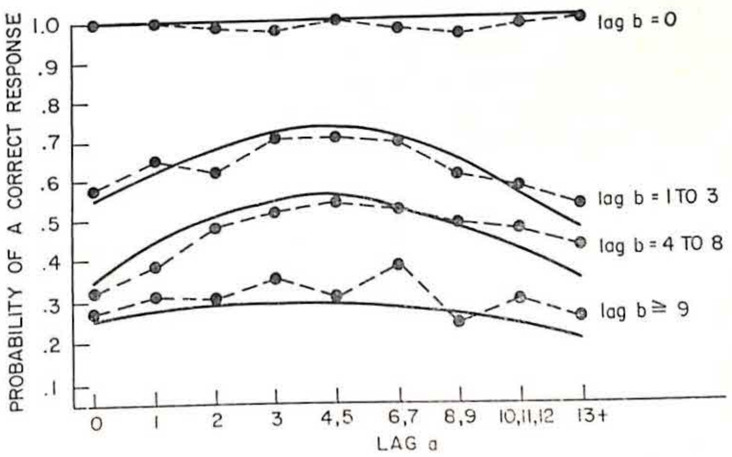


FIGURE 3. Observed and predicted probabilities of a correct response as a function of both lag a and lag b. (See diagram in text.)

for LTS strength to build up. Conversely, a very long lag *a* will result in LTS strength building up to a maximum but then decaying once the item has left the buffer. Hence the probability of a correct response on the second test will be maximal at some intermediate value of lag *a*; namely, at a lag which will give time for LTS strength to build up, but not so much time that excessive decay will occur. For this reason a plot of probability correct on the second test as a function of the lag between the first and second reinforcement should exhibit an inverted U-shape. Fig. 3 is such a plot. The probability correct on the second test is graphed as a function of lag *a*. Four curves are shown for different values of lag *b*. The four curves have not been averaged over all values of lag *b* because we wish to indicate how the U-shaped effect changes with changes in lag *b*. Clearly, when lag *b* is zero, the probability correct is unity and there is no U-shaped effect. Conversely, when lag *b* is very large, the probability correct will tend toward chance regardless of lag *a*, and again the U-shaped effect will disappear.

5. THEORETICAL ANALYSIS

Monte Carlo Methods and Parameter Estimates

The evaluation of the goodness-of-fit of the model involves a number of steps. With the present model and experiment, such desirable statistics as maximum-likelihood estimates of the parameters cannot be calculated; indeed, there is no convenient subset of the exact data sequence which will allow us to calculate ' best ' estimates of the parameters. Our plan of attack will therefore utilize a representative set of results, namely the data of Figs. 2 and 3, to estimate parameter values; these parameter values will then be used to predict a number of additional results to be presented in the next portion of the paper. This section will be concerned with the method by which the model is fitted to the data of Figs. 2 and 3.

The goodness-of-fit criterion to be used is a type of least-square measure that is calculated in the same manner as chi-square. The criterion will be referred to as minimum chi-square with the understanding that this terminology is based upon the method of calculation and not the theoretical distribution of the measure. (The statistics in Figs. 2 and 3 are by no means independent and our goodness-of-fit measure will not be distributed as chi-square.) Given a goodness-of-fit measure, we now must decide how to generate predictions for a given set of parameter values. As noted earlier, the model is too complex for the usual methods to be employed; instead, Monte Carlo techniques are used. The Monte Carlo procedure involved generating pseudo-data on a computer. These data were generated following precisely the rules specified by the model (and the procedure of the experiment). Therefore, these pseudo-data are an example of how real data would look if the model were precisely correct. Whenever an event occurred in the model that was probabilistic, a random number generator was used to determine the occurrence of that event. The same analysis routine that was applied to the real data to generate Figs. 2 and 3 was applied to the pseudo-data to generate theoretical predictions. A large sample of pseudo-data was generated so that the theoretical predictions were stable and accurate. Having generated theoretical predictions for a given set of parameter values, the goodness of fit was then measured by the χ^2 criterion:

$$\chi^2(n, r, \alpha, \theta, \tau) = \sum_i \left\{ \frac{1}{N_i \Pr(C_i)} + \frac{1}{N_i - N_i \Pr(C_i)} \right\} \left\{ N_i \Pr(C_i) - O_i \right\}^2, \qquad (2)$$

where the sum is taken over all data points in Figs. 2 and 3. For the actual data, the observed number of correct responses for the *i*th point is denoted by O_i, and N_i is the total number of responses.

For the Monte Carlo data, $\Pr(C_i)$ denotes the probability of a correct response; of course, the value of $\Pr(C_i)$ depends upon the parameter vector $(r, \alpha, \theta, \tau)$. We denote by n the number of simulated trials upon which $\Pr(C_i)$ is based. As usual, the smaller the χ^2 value, the closer are the predicted and observed data.

In order to determine which set of parameter values provides the best fit according to the χ^2 criterion, a fairly exhaustive search of the parameter space was undertaken. (The grid search procedure is similar to that described in Atkinson & Crothers (1964).) Initially the search was carried out using 4,000 different values for the parameter vector $(r, \alpha, \theta, \tau)$, with a χ^2 value being generated for each vector. For that parameter vector yielding the minimum χ^2, a second, converging, search routine was then begun. These converging searches were continued until the set of parameters generating the minimum χ^2 was determined to two decimal places. One problem that arises when using a Monte Carlo procedure is determining how large a sample of pseudo-data is required to ensure accurate predictions. In this case, an attempt was made to ensure accurate predictions by generating a very large amount of pseudo-data for each set of parameter values: in the initial grid search each set of parameter values were given 200 subjects worth of pseudo-data (each subject consisting of 2200 pseudo-trials). As a check on the stability of the predictions, the pseudo-data were divided in half (100 subjects each) and separate χ^2's computed for each half [$\chi^2(1)$ and $\chi^2(2)$]. The following inequality was then evaluated:

$$0{\cdot}95 \leqslant \frac{\chi^2(1)}{\chi^2(2)} \leqslant 1{\cdot}05. \qquad (3)$$

If this inequality was satisfied then χ^2 was computed for the entire 200 subjects worth of pseudo-data, and the value was assumed to be appropriate for that point in the parameter space. If the inequality failed, then the number of pseudo-trials was doubled and the split-half test again applied. The number of pseudo-trials was increased in this manner until the inequality was satisfied. Furthermore, as the grid search began to converge on the best set of parameter values, the bounds of inequality (3) were narrowed so that even greater accuracy of predicted values would result. When the final set of parameter values was obtained, 12,500 Monte Carlo subjects were generated using those parameter values. In all subsequent discussions, the predicted values are based on the output of this final Monte Carlo run, and it is doubtful that they reflect any fluctuations due to sampling error. The best set of parameter values were: $r=3$; $\alpha=0{\cdot}65$; $\theta=1{\cdot}25$; $\tau=0{\cdot}82$.

These parameter values are in reasonable accord with those found in previous experiments using the same model (Atkinson & Shiffrin, 1968). The buffer size of $r=3$ may seem small at first glance but r has been estimated to be 2 or 3 in each of the experiments of the present type. It is not hard to see that a small value of r is to be expected: an r of 3 indicates that the subject is simultaneously rehearsing six numbers and three letters, a fairly difficult task considering interruptions for tests and changes in rehearsal following study periods. The value of τ is fairly low, considering that this decay factor is applied on each trial; one explanation for LTS strength being reduced so quickly would hold that there is a great deal of retroactive interference in this situation. If so, then there should be evidence of this in the data. Evidence along these lines will be presented shortly.

The predictions from the theory are shown as the smooth curves in Figs. 2 and 3. It should be evident that the predicted values are quite close to the observed ones. Note also that the seven curves in the two figures are fitted simultaneously with the same four parameter values. The fact that the spacing of the curves is accurately predicted is particularly interesting.

Some Further Predictions

A number of statistics that were not used in estimating parameters are now examined. These statistics test specific predictions of the model, predictions that were in some cases contrary to the authors' *a priori* intuitions. Most of the machinery of the model, and its most novel features, lie in the rehearsal scheme called the buffer. Fortunately, the very specific assumptions made concerning the working of the buffer lead to clear-cut predictions to be searched for in the data. Consider, for example, the hypotheses concerning *O*- and *M*-items. The model predicts that the kind of items intervening between study and test will influence the probability correct at test. For one thing, the more *N*-items intervening between study and test, the less the probability correct, because only *N*-items can eliminate an item from the buffer. Similarly, the more intervening *O*-items, the greater the probability correct, because *O*-items cannot knock the studied item from the buffer. Although *O*-items and *N*-items are not directly identifiable in the data, the probability of their occurrence can be maximized by selecting appropriate event sequences. Thus in Fig. 4 the 'all-same' and 'all-different' curves are plotted. For the all-same curve, the probability of a correct response is computed as a function of the lag, when all the intervening items between study and test utilize the same stimulus. There are three such curves, depending upon whether the studied item had been given

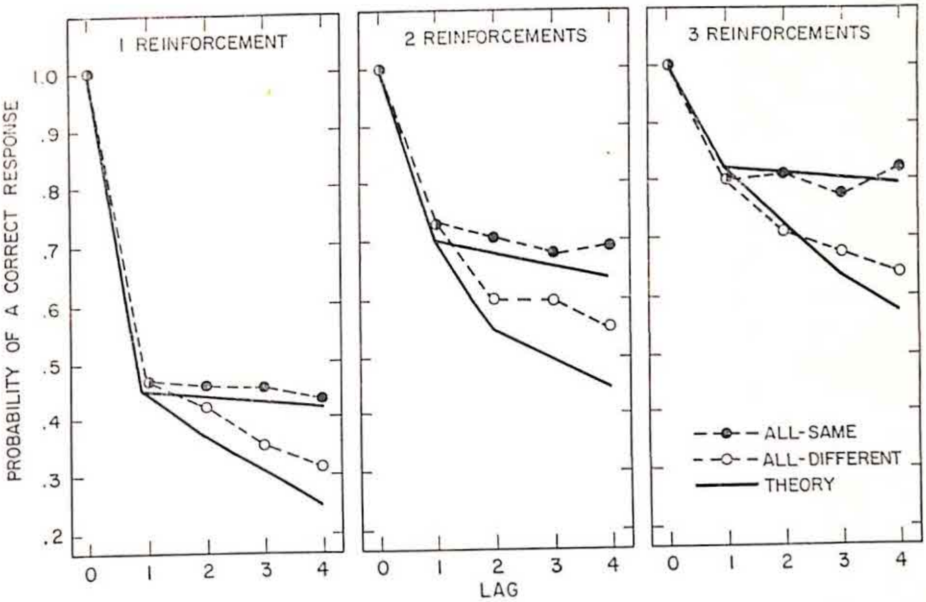


FIGURE 4. Observed and predicted lag curves for the all-same and all-different conditions.

its first, second or third reinforcement. The model predicts that once any of the intervening items enters the buffer, every succeeding item will be an *O*-item and hence will also enter the buffer (because all these intervening items have the same stimulus). Thus the all-same curve should decrease less (as a function of lag) than the unconditional lag curves presented in Fig. 2. For the all-different curves, on the other hand, the probability of a correct response is computed for instances when every intervening item between study and test utilizes a different stimulus. In this case, the number of intervening *N*-items tends to be maximized, and hence the probability that the studied item will be knocked out of the buffer tends to be maximized. Therefore, the all-different curves should decrease faster than the unconditional lag curves. It will be seen in Fig. 4 that the all-same and all-different curves conform to these predictions. The solid lines in the figure represent predictions from the model using the parameter values estimated in the previous section; the correspondence of data and theory appears to be reasonably close.

Next consider the factors determining the probability that a presented item will enter the buffer. Most important is the probability that the presented item is an *O*-item, since an *O*-item's stimulus is already in the buffer, and thus every *O*-item enters the buffer. Because a high probability of entering the buffer implies a high probability of a correct response at test, it should be possible to manipulate the probability correct by manipulating the probability that a presented item is an *O*-item. In Fig. 5 this has been done. Consider a sequence of consecutive trials all utilizing the same stimulus, but with the last item in the sequence being given its first reinforcement (thus its response will be different from that of the immediately preceding item). Once any item in this sequence enters the buffer, every following item will do so also; thus the longer the sequence, the greater the probability that the last item in the sequence will enter the buffer. Fig. 5 plots the probability of a correct response as a function of the length of the sequence of preceding items all utilizing the same stimulus. The curve plotted is *not* a lag curve; the probability correct is pooled for over all lags at which the eventual test occurred. The theoretical predictions are generated from the previously estimated parameter values, and again there is a good correspondence between theory and data. The effect found here is particularly important because it emphasizes the dichotomy between short- and long-term processes. A traditional interference theory would seem to predict just the opposite effect from that found, in that ' proactive interference ' should increase as the length of the preceding sequence increases. Nevertheless, indications of typical interference effects will be seen for long lags, where LTS, and not the buffer, is playing a predominant role.

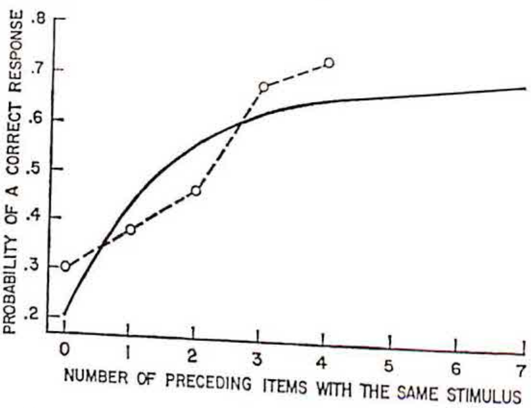
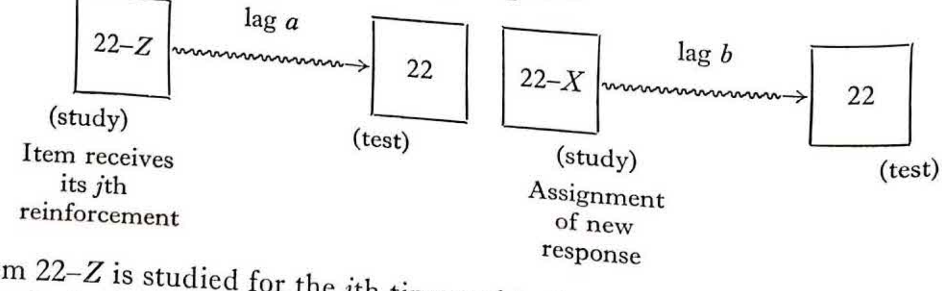


FIGURE 5. Observed and predicted probabilities of a correct response as a function of the number of consecutive preceding items using the same stimulus.

In order to extend and verify the results of Fig. 5, consideration is now given to the effect of the lag preceding an item's presentation for study. To make matters clear, consider the following diagram:


22–*Z*
lag *a*
22
22–*X*
lag *b*
22
(study)
Item receives its *j*th reinforcement
(test)
(study)
Assignment of new response
(test)

Item 22–*Z* is studied for the *j*th time and is then tested at lag *a*; on this trial 22 is paired with a new response *X*, and tested next at lag *b*. According to the theory, the shorter lag *a*, the better performance should be when the item is tested after lag *b*. This prediction is based on the fact that the more recently a stimulus had appeared, the more likely that it was still in the buffer when the next item using it was presented for study. If the stimulus was in the buffer, then the item using it is an *O*-item and automatically enters the buffer. In the present analysis, we examine this effect for three conditions, i.e. the preceding item using the stimulus in question could have just received its first, second or third reinforcement. Fig. 6 presents the appropriate data. In terms of the above diagram, what is plotted is the value of lag *a* on the abscissa versus the probability of a correct response averaged over all values of lag *b* on the ordinate. There is a separate curve for $j=1$, 2 and 3. Note that the results of Fig. 5 are confirmed by those of Fig. 6; again interference theory would appear to predict an effect opposite to that found.

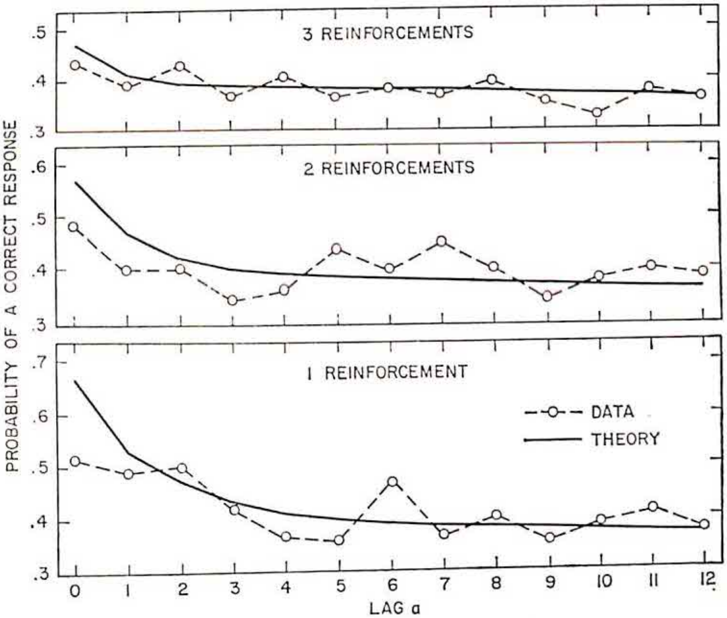


FIGURE 6. Observed and predicted probabilities of a correct response as a function of lag *a*. (See diagram in text.)

The predicted curves are based upon the previous parameter estimates. The predictions and observations coincide fairly well, but the effect is not as dramatic as one might hope. One problem is that the predicted decrease is not very large. Considerably stronger effects may be expected if each curve is separated into two components: one where the preceding item was correct at test and the other where the preceding item was not correct. In theory the decrease predicted in Fig. 6 is due to a lessened probability of the relevant stimulus being in the buffer as lag *a* increases. Since an item in the buffer is always responded to correctly, an analysis made conditional upon correct responses or errors (the centre test in the above diagram) should magnify the effect. To be precise, the decrease will be accentuated for the curve conditional upon correct responses, whereas no decrease at all is predicted for the curve conditional upon errors. If an error is made, the relevant stimulus cannot be in the buffer and hence the new item enters the buffer with probability α, which is independent of lag a. Fig. 7 presents the conditional curves and the predictions. The decreasing effect is fairly evident for the correct curves, whereas the error curves, as predicted, are quite flat over lags. Conceivably one might argue that the effects are due to item selection with correct responses indicating easier stimuli and incorrect responses indicating more difficult ones. However, it is difficult to imagine how item selection could explain the eventual crossing of the correct and error curves found in each of the three diagrams. (Undoubtedly there are some selection effects in the data of Fig. 7, but their magnitude is difficult to determine. Thus, these data should be regarded with some caution.) Indeed, the model does not explain the crossover. The model predicts that the two curves should meet. The model is in error at this point because it has not been extended to include negative transfer effects. Such an extension would not be difficult to implement, however. An item responded to correctly at a long lag probably has a strong LTS trace. This strong trace would then interfere with the LTS trace of the new item which, of course, uses the same stimulus. Thus the data of Fig. 7 appear to imply opposite effects at

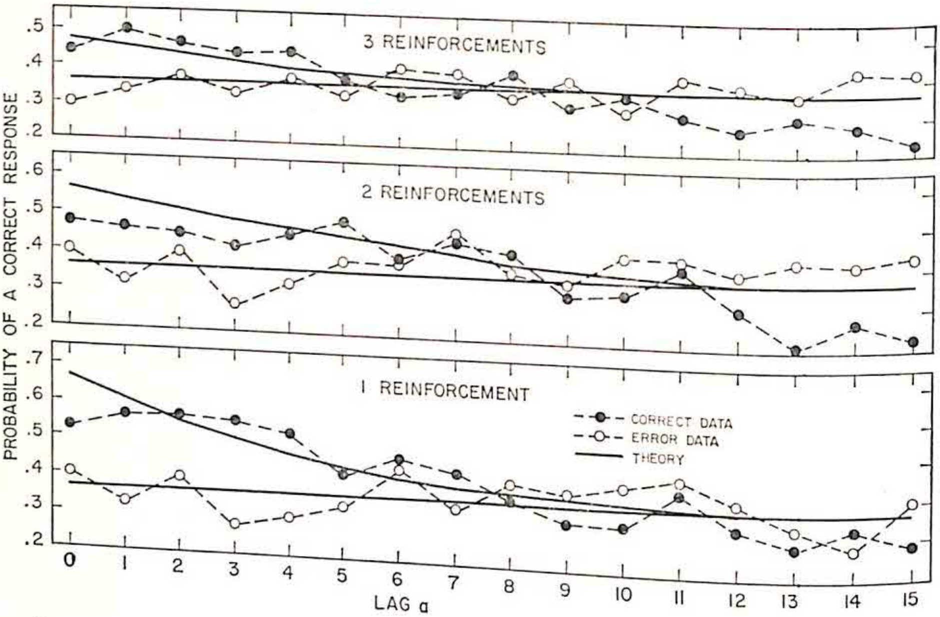
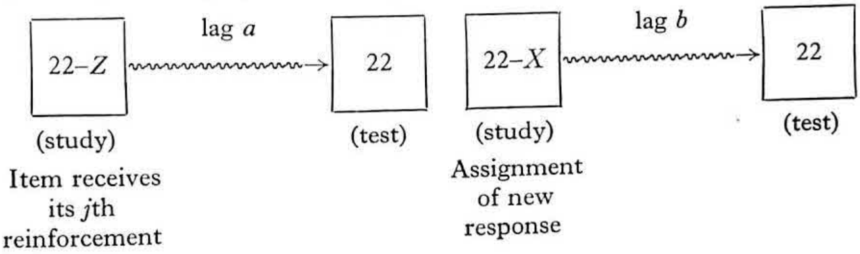


FIGURE 7. Observed and predicted probabilities of a correct response as a function of lag a conditional upon correct and incorrect responses. (See diagram in text.)

short and long lags, with the effect at long lags (when LTS rather than the buffer may be assumed to be the predominating influence) exhibiting the result expected from interference theories.

Figs. 6 and 7 both show that the predicted decrease becomes smaller as the number of reinforcements increases. The fact that the data seem to support this prediction helps confirm certain of the buffer-replacement assumptions. The decreasing effect as reinforcements increase is predicted because the probability of entering the buffer is reduced for an item receiving its third reinforcement; this occurs as a consequence of the assumption that an item retrieved from LTS on the test phase of the trial is not entered into the buffer on the study phase of that trial. Thus as reinforcements increase, the probability of being in the buffer as a result of short lag *a* is partially counterbalanced.

By and large, the data and predictions to this point may be considered to provide fairly strong support for the details of the model. The feature that has been left out of the model is that of LTS response competition, or negative transfer. The model fails to take account of this effect because it ignores residual information in LTS from previous items using the same stimulus. This lack is most clearly indicated by the occurrence of intrusion errors; particularly, errors that were correct responses on the preceding occurrence of that stimulus. For example, consider the following sequence:


lag *a*
22–*Z*
22
22–*X*
lag *b*
22
(study)
Item receives its *j*th reinforcement
(test)
(study)
Assignment of new response
(test)

Item 22–*Z* is studied for the *j*th time and then tested at lag *a*. On this trial 22 is paired with a new response *X* and next tested at lag *b*. By an intrusion error we mean the occurrence of response *Z* when 22 is tested at the far right of the diagram. The model predicts that these intrusion errors will be at a chance level (1/25) and independent of lag and number of reinforcements. In fact, these predictions fail. Fig. 8 presents the probability of intrusion errors as a function of lag *b*, where the data are pooled for all values of lag *a*. Three curves are plotted for $j=1$, 2 and 3. If it is assumed that the probability of giving the previously correct response as an intrusion error is some function of that previous item's current LTS strength, then the pattern of results in Fig. 8 follows naturally. For example, the more reinforcements of the previous item, the greater its LTS strength, and the greater the probability of its response being given as an intrusion error. Thus this failure of the model is not very distressing, rather it was expected. The model could be extended in a number of obvious ways to take account of competing LTS traces without appreciably changing the accuracy of the predictions already considered.

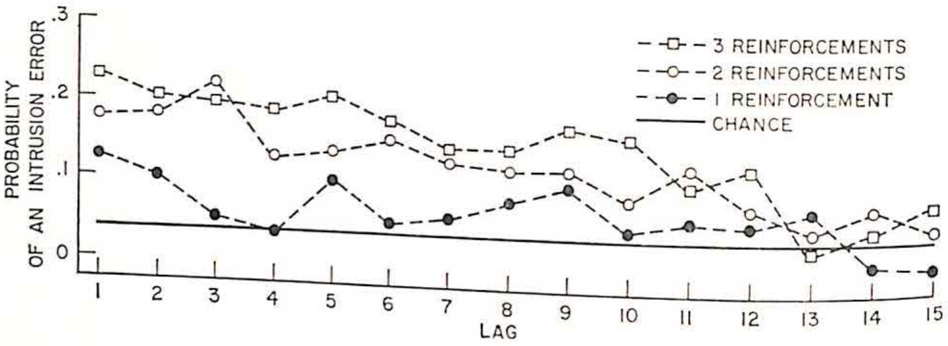


FIGURE 8. Probability of an intrusion error as a function of the prior number of reinforcements.

The major emphasis in this paper has been on rehearsal processes conceptualized in the framework of a buffer mechanism. Because of this emphasis upon short-term processes, the experiment should not be considered as a strong bridge to the usual paired-associate learning situation. However, a number of long-term effects (such as intrusion errors and interference caused by previously learned items on new items using the same stimulus) demonstrate that LTS mechanisms play an important role in this study. It is undoubtedly true that modifications of the theory are required before it can be applied to the typical paired-associate learning experiment. For example, it would be necessary to provide more structure for the workings of LTS. Several possible forms for such structure have been suggested by Atkinson & Shiffrin (1968). These new considerations have yet to be experimentally investigated, but it appears that the theoretical gap between 'memory' and 'paired-associate learning' tasks may be bridged without too much difficulty in the near future.

ACKNOWLEDGEMENTS

This research was supported by the National Aeronautics and Space Administration, Grant No. NGR-05-020-244. The authors wish to thank Susan Shiffrin for collecting the data reported in this paper.

REFERENCES

ATKINSON, R. C., BRELSFORD, J. W. & SHIFFRIN, R. M. (1967). Multi-process models for memory with applications to a continuous presentation task. *J. math. Psychol.* **4**, 277–311.

ATKINSON, R. C. & CROTHERS, E. J. (1964). A comparison of paired-associate learning models having different acquisition and retention axioms. *J. math. Psychol.* **1**, 285–315.

ATKINSON, R. C. & SHIFFRIN, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 2. New York: Academic Press.

YNTEMA, D. B. & MUESER, G. A. (1960). Remembering the present state of a number of variables. *J. exp. Psychol.* **60,** 18–22.

YNTEMA, D. B. & MUESER, G. A. (1962). Keeping track of variables that have few or many states. *J. exp. Psychol.* **63,** 391–395.

Vol. 21
Part 1
21–33

The British Journal of Mathematical and
Statistical Psychology

May
1968

A RETEST OF THE RESPONSE-BIAS EXPLANATION OF THE WORD-FREQUENCY EFFECT

By John Morton

Medical Research Council Applied Psychology Research Unit, Cambridge

This paper investigates whether the relation between the frequency of occurrence of a word and its sensory threshold can properly be ascribed to response bias. Seven models for the recognition process are tested against data of Brown & Rubenstein (1961). The models are of two kinds: probabilistic single-threshold models and information-processing models. It is concluded that the data can only be accounted for within the former class of models by assuming a differential effect of the stimulus as well as response bias. For the class of information-processing models, the data require the assumptions that there is equal sensitivity for words of all frequencies and a lower criterion for more common words. The relation between the two acceptable models makes it apparent that the notions of 'stimulus effect' and 'response effect' are by no means as clear as has previously been thought.

1. Introduction

When individual words are presented acoustically in noise or exposed visually for a brief period, the probability of an observer identifying them correctly increases with the frequency of occurrence of the stimulus word in the language. It is at present a matter of dispute as to whether this result is due to stimulus effects or to response bias (Broadbent, 1967).

Ultimately it will be shown in the present paper that the rigid dichotomy between stimulus and response factors is, in part, a semantic artifact, for it is possible to devise information-processing models to which neither of these terms can be singly and unambiguously applied. The dichotomy, as it is usually employed, may be restated in the following way.

1. It is possible that high-frequency words are more distinctive than words of lower frequency either by virtue of their 'pure' stimulus properties or because the recognition system is organized in such a way as to enable common stimuli to be processed more readily than less common stimuli.

2. The response-bias explanation of the word-frequency effect would deny the above possibilities, attributing the results entirely to the observer's tendency to use high-frequency words more often. Thus in the absence of any stimulus information, or if such information were minimal and insufficient for a clear 'percept', the observer would respond according to his predispositions, that is 'guess', and would respond preferentially with high-frequency words.

Brown & Rubenstein (1961) claim that when response bias is controlled the word-frequency effect is eliminated. To eliminate the possibility of response bias affecting the data they introduce the notion of frequency 'intervals',

Bureau of Ednl.
S. C. E. R. T.
Date

interpreting the response-bias explanation as implying ‘ that groups of words which differ in word-frequency are random samples from the same population of acoustical or visual parameters ’. This assumption, termed the Principle of Acoustical Equivalence, has the force that no information as to the frequency interval of the word can be gained from the stimulus and any word is likely to be confused on the basis of its stimulus properties with a word from any of the frequency intervals.

They divided up the 6500 monosyllabic content words into 13 intervals of 500 words each by their frequency of occurrence. Six subjects were then presented with 1300 words (100 from each interval in random order) at two speech-to-noise ratios (S/N). Three numbers were determined for each frequency interval *i*:

c_i the number of correct responses to stimuli in the *i*th interval;

s_i the number of responses whose frequency class corresponded to that of the stimulus (this would include both correct and incorrect responses);

r_i the total number of responses in the *i*th interval to all stimuli.

TABLE 1. RAW DATA FROM BROWN & RUBENSTEIN

(Maximum values of c_i and s_i are 600. Maximum value of r_i is 7800.)

	S/N=0 db			S/N=10 db		
i	c_i	s_i	r_i	c_i	s_i	r_i
1	21	34	77	325	332	367
2	23	28	81	313	318	361
3	55	68	135	390	394	448
4	29	45	161	344	355	437
5	29	39	159	357	371	450
6	47	62	227	415	424	534
7	55	76	255	414	419	522
8	78	114	409	445	460	617
9	105	145	608	428	447	618
10	101	155	674	427	442	657
11	125	183	924	473	479	742
12	129	213	1301	415	433	739
13	242	388	2428	487	527	883
0			361			425

The values of these three variables at the two S/N are given in Table 1. Brown & Rubenstein derive equations to test the response-bias explanation; these equations do not, however, follow from the premises and are discussed briefly in the Appendix. In addition, they base some conclusions on the linearity of the relation between c_i and s_i. Since the latter is included in the former, such a procedure seems invalid: for the higher S/N it virtually involves plotting c_i against itself. In the present paper a different datum is used: e_i, the number of error responses which are in the correct frequency interval. This equals $s_i - c_i$. The values of e_i at the higher S/N are too small, and are too irregular with respect to i to have value in a graph. The relation between e_i and c_i for the lower S/N is depicted in Fig. 1. The data are plotted in terms of proportions, and the variables will be treated in this way in the body of the paper.

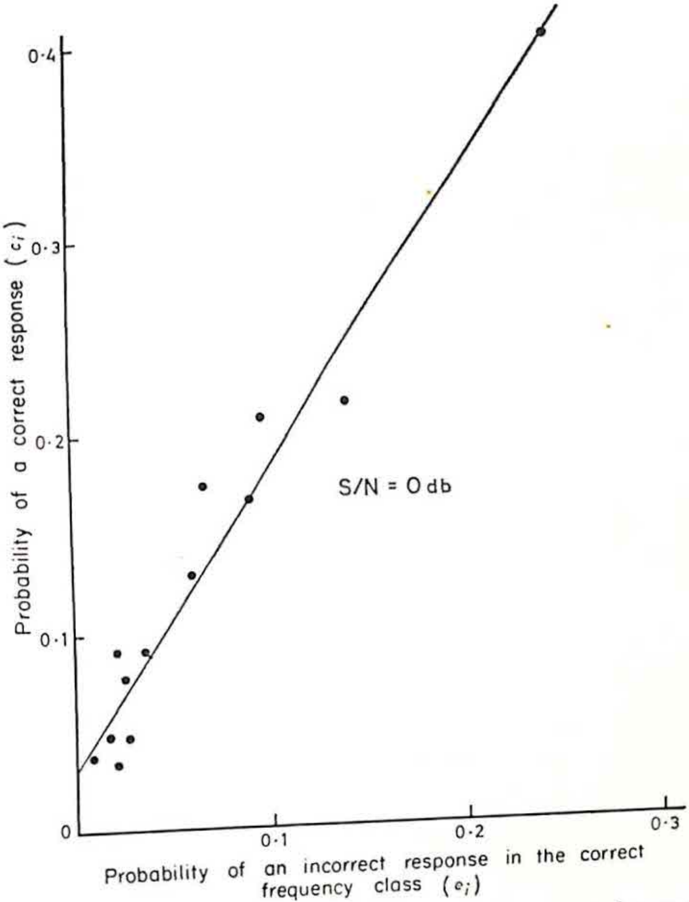


FIGURE 1. Performance in the recognition of visually presented words of different frequencies of occurrence. (Data from Brown & Rubenstein, 1961.)

The line drawn is the best-fitting straight line:

$$c_i = 0{\cdot}033 + 1{\cdot}526\, e_i;$$

and the best-fitting quadratic is:

$$c_i = 0{\cdot}024 + 1{\cdot}755\, e_i - 1{\cdot}356\, e_i^2.$$

A further datum used below is: p_i, the total number or proportion of error responses in the ith interval to *all stimuli*. This equals $r_i - c_i$ in terms of numbers and $r_i - c_i/n$ in terms of proportions. A plot of c_i against p_i is given in Fig. 2 for both S/N values.

Brown & Rubenstein's data will be used to test a number of models for word recognition. Some of these models have been explicitly or implicitly used by other authors (see Broadbent, 1967, for references). Theory E, however, has not previously been examined in any detail.

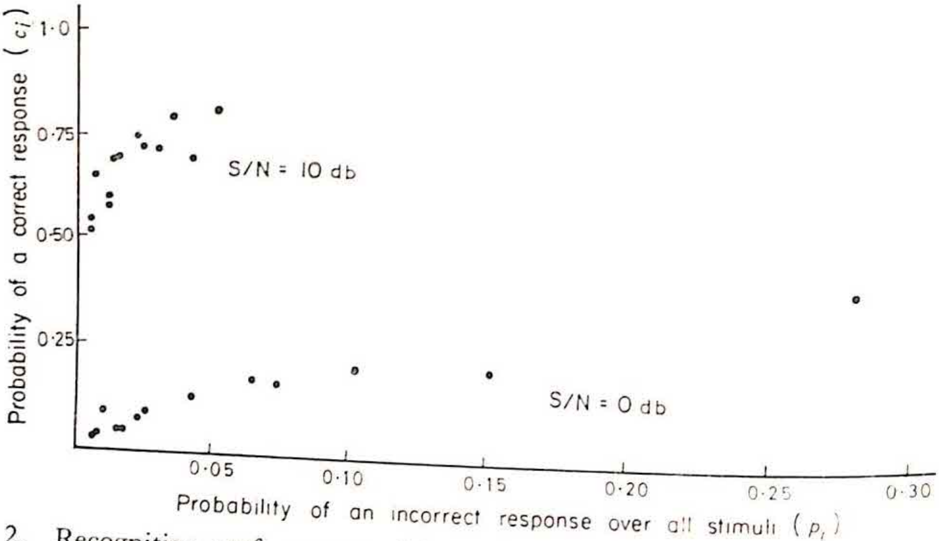


FIGURE 2. Recognition performance with words of different frequencies of occurrence. (Data from Brown & Rubenstein, 1961.)

2. SINGLE-THRESHOLD THEORIES

The first group of models examined are single-threshold models. According to these, the subject guesses in the absence of a percept.

Theory A. Either a word is recognized or it is not. If it is not, then the observer guesses according to his dispositions which are independent of the stimulus. By the response-bias theory, the probability of a word being recognized, x, increases with S/N but is independent of the interval i. If the stimulus is not recognized, then the observer guesses, the proportion of guesses in interval i being g_i. Thus with n intervals and M words in each interval, given a stimulus in the ith interval, it follows that:

$$s_i = x + (1-x)g_i, \tag{1}$$

$$c_i = x + \frac{1}{M}(1-x)g_i, \tag{2}$$

$$e_i = \left(1 - \frac{1}{M}\right)(1-x)g_i. \tag{3}$$

Therefore, eliminating g_i from eqns. (2) and (3):

$$c_i = \frac{1}{M-1}e_i + x. \tag{4}$$

Also, given stimuli equiprobable in i,

$$r_i = (x/n) + (1-x)g_i \tag{5}$$

and

$$p_i = \left(1 - \frac{1}{nM}\right)(1-x)g_i. \tag{6}$$

Therefore,

$$c_i = [n/(nM-1)]p_i + x. \tag{7}$$

Eqn. (4) should correspond to Fig. 1; however, the slope of the data line is too steep (since in the equation $M = 500$ and $n = 13$). Likewise, the slopes predicted for the relations depicted in Fig. 2 from eqn. (7) are too steep, even assuming linearity. The simple theory is therefore incorrect.

Theory B. If the stimulus contains information relating to its own class then the proportion of responses in the correct class increases. This affects the second term on the right-hand side of eqns. (1), (2) and (3), but leaves eqn. (4) and the slope of eqn. (7) unchanged. The modification of Theory A must therefore also be abandoned.

Theory C. It is possible, and intuitively obvious, that on a number of occasions, while insufficient stimulus information is available to enable the observer to recognize the stimulus, there will, however, be enough to enable him to make a choice from a restricted set of alternatives. Thus if the initial phoneme /t/ is heard clearly, only words beginning with ‘ t ’ will be considered by the observer. By the theory of acoustical equivalence, however, this restricted set of alternatives will be equally distributed among the separate response intervals. The forms of eqns. (1) and (5) will therefore remain identical. Eqn. (2) will become:

$$c_i = x + \frac{1}{m}(1-x)g_i, \tag{2 a}$$

where m is the number of words in the reduced set in each interval. Thus eqn. (4) becomes:

$$c_i = \frac{1}{m-1}e_i + x, \tag{4 a}$$

and eqn. (7) becomes:

$$c_i = [n/(nm-1)]p_i + x. \tag{7 a}$$

In eqns. (4 a) and (7 a), m has a lower limit of unity, since it refers to the number of possible words *in the correct class* in the original definition of c_i; the occurrence of m in the last two equations is determined entirely by this definition. Clearly, in the limit the strict definition of the principle of acoustical equivalence breaks down, for it is possible that only two words are possible responses, one of which will be the correct one. The modified principle would state that the other word is on average equally likely to come from any *other* interval. This limiting modification does not, however, affect the equations.

Given that $m \geqslant 1$, the slope of the function defined by eqn. (7 *a*) has a maximum of $13/12 = 1{\cdot}083$. If the (unlikely) assumption is made that the data in Fig. 2 are indeed linear, the equations of the best-fitting straight lines are:

$$c_i = 5{\cdot}09\, p_i + 0{\cdot}56 \quad (\text{for S/N} = 10 \text{ db}),$$

$$c_i = 1{\cdot}29\, p_i + 0{\cdot}05 \quad (\text{for S/N} = 0 \text{ db}).$$

Both these lines have slopes in excess of the permitted maximum and model C is thus untenable on the basis of Brown & Rubenstein's data.

Theory D. A combination of models B and C still yields eqn. (4 *a*) and a form of eqn. (7 *a*) with the same slope. This theory cannot be accepted either. It would appear, then, that threshold models which deny the stimulus any role in the word-frequency effect cannot account for the data.

Theory E. By denying the common assumption of the above group of theories, the primary data can be accounted for quite satisfactorily. It will be accepted, then, that the proportion of correct responses resulting from the stimulus alone, x_i, is a function of the interval i as well as of S/N, initially assuming that the stimulus contains no information as to its interval and that the stimulus is either recognized or guessed at. The following equations are then obtained:

$$c_i = x_i + \frac{1}{M}(1 - x_i)g_i, \tag{8}$$

$$e_i = \left(1 - \frac{1}{M}\right)(1 - x_i)g_i, \tag{9}$$

$$p_i = \left(1 - \frac{1}{nM}\right)(1 - X)g_i, \tag{10}$$

where $X = \sum_{i=1}^{n} (x_i/n)$. From these equations the following can be derived:

$$c_i = \frac{1}{M+1} e_i + x_i, \tag{11}$$

$$c_i = p_i \frac{n(1 - x_i)}{nM(1 - X) - (1 - x_i)}. \tag{12}$$

Since x_i is now a function of i, eqn. (8) can now represent Fig. 1 without the assumption that M is small, and the slope limitations on the form of Fig. 2 no longer apply. Equally it is not necessary for either function to be linear. There are then no gross objections to this model.

If x_i is eliminated from the modified eqns. (2) and (3), then

$$c_i = 1 - e_i \frac{M-g}{g(M-1)}. \tag{13}$$

Thus

$$g_i = \frac{e_i M}{e_i + (1-c_i)(M-1)}. \tag{14}$$

If M is large, we may simplify and rearrange this equation to obtain:

$$g_i = e_i/(1-c_i). \tag{14 a}$$

Values of g_i were calculated from this equation and are tabulated in Table 2 (*a*). If the model were accurate, the sum of the values of g_i plus g_0, the probability of producing a response outside the classification system (i.e. blanks, partial responses and polysyllabic words) should be unity. This value is, in fact, exceeded at both S/N. This is not a consequence of the approximation, since the effect of reducing the value of M in eqn. (11) is to increase the calculated values of g for given values of c_i and e_i. The values of g being too high could result from the values of either c_i or e_i being too high compared with the values of the other. One possible explanation of this is that words do carry some information concerning their frequency of occurrence which would have the effect of increasing e_i relative to c_i. A second anomaly in Table 2 (*a*) is that values of g for high frequency words ($i = 10$ to 13) are proportionally lower at the high S/N. These two points will be discussed below.

TABLE 2. VALUES FOR g_i CALCULATED FROM (*a*) EQN. (14 *a*), (*b*) EQN. (15)

	(*a*) $g_i = e_i/(1-c_i)$		(*b*) $g_i = p_i/(1-c_i)$	
i	S/N = 0	S/N = 10	S/N = 0	S/N = 10
1	0·023	0·026	0·008	0·016
2	0·009	0·017	0·009	0·019
3	0·024	0·019	0·012	0·023
4	0·028	0·043	0·020	0·036
5	0·018	0·058	0·019	0·036
6	0·027	0·049	0·027	0·046
7	0·038	0·027	0·030	0·042
8	0·069	0·097	0·049	0·067
9	0·081	0·111	0·074	0·074
10	0·108	0·087	0·084	0·089
11	0·122	0·047	0·118	0·104
12	0·178	0·097	0·173	0·126
13	0·408	0·354	0·323	0·154
Sum	1·133	1·032	0·946	0·834
Others	0·054	0·166	0·054	0·166
Sum	1·187	1·198	1·000	1·000

If M is large, then $c_i \doteqdot x_i$, and $p_i \doteqdot (1-X)g_i$ from eqns. (8) and (10). The value of X can then be estimated from the average number correct and we may write:

$$g_i = p_i/(1-C), \tag{15}$$

where $C = \sum c_i/n$. These estimates of g are given in Table 2 (*b*). The form of eqn. (15) ensures that these values sum to unity. As with Table 2 (*a*), the values of g_i for higher values of i are consistently lower at the higher S/N.

3. INFORMATION-PROCESSING MODELS

The above models have as their basis the probability of certain events taking place. If instead we regard word recognition in terms of information processing we obtain different equations, use different parameters, and discover different implications for the notions of stimulus effect and response effect.

The basic unit to be considered is one which accumulates evidence from the stimulus. When more than a certain amount of evidence is available, the word is recognized. On the assumption that all the units are inherently noisy, we can treat them in terms of signal-detection theory (Green & Swets, 1966). The two relevant parameters are then the amount of stimulus information available, d', and the thresholds or criteria of the recognition units, β.

Theory F. A model which was purely stimulus based, in the sense described in the introduction, would say that the value of d' was a function of the frequency of occurrence of the word as well as of S/N. 'Active' theories of speech perception (Halle & Stevens, 1962; Liberman, Cooper, Harris & MacNeilage, 1963) would have to predict such a result. It is uncertain as to whether these theories would have the criteria dependent on frequency.

Theory G. An alternative model is one in which d' remains constant for all words but the value of β changes with frequency. In such a model, less stimulus evidence is required for a high-frequency response.

By analysing the data under discussion in terms of response strengths (Luce, 1959), a form of analysis roughly equivalent to signal-detection theory but considerably easier to handle mathematically, it will be shown that model F cannot handle the data, whereas model G can.

A matrix can be constructed showing, for a stimulus of a particular frequency interval, the relative response strengths for the correct response, for incorrect responses of all intervals, and for other responses. This matrix is given in Table 3. V_i is the relative response strength for a word in the absence of a stimulus and is related to the criteria; α is a multiplicative factor representing the stimulus strength which is related to d'. The value of α represents a contribution to the response strength of the correct response compared with the *average* of other words in the interval. This model also assumes the acoustic equivalence of intervals. The probability of a particular response is then given by the response strengths for that response divided by the sum of all response strengths for that stimulus. Given a stimulus in interval i, the proportion of correct responses, c_i, is given by:

$$c_i = \frac{\alpha V_i}{(\alpha-1)V_i+(MZ+1)}, \tag{16}$$

where $Z=\sum V_j$ $(j=1, \ldots, i, \ldots, n)$, which we may more simply write as

$$c_i=\alpha V_i/T_i. \tag{16 a}$$

The proportion of incorrect responses in the correct frequency interval, e_i, is given by:

$$e_i=[(M-1)V_i]/T_i. \tag{17}$$

From these two equations the following relation is obtained:

$$c_i=e_i[\alpha/(M-1)]. \tag{18}$$

TABLE 3. RESPONSE STRENGTHS

Stimulus interval	Correct	Incorrect responses by interval					Other responses	Total
		1	…	i	…	n		
1	$\alpha_1 V_1$	$(M-1)V_1$	…	MV_i	…	MV_n	1	T_1
i	$\alpha_i V_i$	MV_1	…	$(M-1)V_i$	…	MV_n	1	T_i
n	$\alpha_n V_n$	MV_1	…	MV_i	…	$(M-1)V_n$	1	T_n

Fig. 1 shows this prediction to be largely upheld. The possible negative curvature and the non-zero intercept could be interpreted simply as due to the subjects recognizing words which they had never before seen and so could not produce as incorrect responses; they would effectively be nonsense syllables. Any recognition model would have to treat such words as exceptional cases. It might be noted that such an explanation would serve to account for the non-unit sum of g_i in Table 2 (*a*).

If model F were correct, Fig. 1 would be positively accelerating since α would increase as i (and so c_i and e_i) increased. This prediction is emphatically rejected by the data.

4. A COMPARISON OF MODELS E AND G

The conclusion that α is constant across i may seem to be in conflict with the result from model E where it was seen that x_i had to vary with i. However, while α can only be interpreted in terms of the amount of information available from the stimulus, x_i merely specifies the proportion of correct responses which result from the stimulus. From eqns. (8), (9), (16) and (17) the relations among the parameters in the two models are given by:

$$x_i = \frac{(\alpha-1)V_i}{(\alpha-1)V_i+(MZ+1)}, \tag{19}$$

$$g_i = MV_i/(MZ+1). \tag{20}$$

Thus x_i represents a combination of both stimulus and response strength factors and g_i is proportional to V_i, being specifically the probability of making a response in the ith interval in the absence of a stimulus, i.e. when $\alpha=1$.

If eqn. (20) is summed over i we obtain:

$$\sum_i g_i = MZ/(MZ+1). \tag{21}$$

However, it has already been observed that the sum of the obtained values of g is greater than unity. Thus theory G also requires some modification to account completely for the data if the explanation following eqn. (18) is not acceptable. Two possibilities remain to be examined: that there is some information in the stimulus concerning its interval, and that an incompletely recognized stimulus results in a reduced set of alternatives.

For theory E, this would give equations such as:

$$c_i = x_i + \frac{1}{m}(1-x_i)[y_i+(1-y_i)g_i], \tag{22}$$

where m is the size of the reduced set and y_i is the proportion of occasions when a stimulus in the ith interval is not recognized but the correct interval is obtained from the stimulus. It is reasonable to assume that both m and y_i are functions of the S/N. From such equations we can derive values for g_i from the relation:

$$g_i = \frac{me_i - y_i[(1-c_i)(m-1)+e_i]}{(1-y_i)[(1-c_i)(m-1)+e_i]}. \tag{23}$$

As m is reduced, the values of g_i increase, and as y increases g decreases. When Brown & Rubenstein's data are entered in this equation, it turns out that there are no values of y and m such that $\sum g < 1$ for a constant y, which also give positive values for all g_i; this comes about since, if y is increased to reduce the values of g_i, the right-hand negative side of the numerator exceeds the positive part for low values of i. Therefore, within this model, the value of y must be a function of i; that is, an unrecognized high-frequency word gives more information about its frequency interval than does a low-frequency word. This modification does not affect the acceptability of theories C and D since eqn. (7 a) is unaffected.

With theory G, interval information would be expressed by a multiplicative factor on the response strengths for all responses in the correct interval. This factor will be called δ. A relationship between c_i and e_i is obtained which is unchanged by δ but does include the size of the reduced set of alternatives m. This is

$$c_i = e_i[\alpha/(m-1)]. \tag{24}$$

The equivalent of eqn. (20) is the following:

$$\frac{mV}{mZ+1} = \frac{me_i}{(m-1)[\delta(1-e_i)]+me_i}. \tag{25}$$

This equation must produce positive values for V_i. As δ is increased, the value of the function falls, and as m decreases the value rises: this is analogous to the properties of eqn. (23). For any m, δ can be adjusted to fulfil the restrictions on the sum of the left-hand side of the above equation (cf. eqn. (21)) without having to postulate that δ varies with i.

By either of theories E and G, Brown & Rubenstein's data can only be accounted for by assuming that the stimulus contains information about its interval. With theory E, it is in addition necessary to propose that such information is greater with high- than with low-frequency words. Such a proposition could work in practice through partial information and as such would be associated with the size of the reduced set which was considered above. For example, if all that was recognized clearly was the letter 'z', then the set of possible responses would be relatively biased in favour of low-frequency words. On the other hand, to give a hypothetical example, since the necessary statistics are not available: if the letter sequence '–the–' were recognized clearly, the set of alternatives could be biased in favour of high-frequency words.

The large change in the relative sizes of the entries in Table 2 at the higher S/N, which affects g_i in theory E and V_i in theory G, requires a further construct. The observation amounts to saying that as more stimulus information becomes available, proportionately fewer high-frequency words are produced erroneously. This could be accounted for in both the theories by postulating that whereas the amount of interval information given by a stimulus is greater for high- than low-frequency words, the difference becomes less at higher S/N ratios. This would apply to either y_i or δ in the two theories.

An alternative way of explaining this finding would be to assume that the recognition-system, while requiring less positive stimulus information for a high-frequency response to be made, also requires less negative stimulus information for such a response to be rejected. For lower-frequency words the system would be relatively indifferent to negative information, being only concerned to accept positive evidence. Such a proposition would require a far more detailed specification of the recognition system to be tested.

5. CONCLUSIONS

It has been shown that while a simple guessing model cannot account for the word-frequency effect, a more complex one can fit the data. The disadvantage of this theory compared with the one derived from signal-detection theory is that there are two monotonically related variables in the theory, x_i and g_i. Model G, on the other hand, predicts the relation between c_i and e_i from its initial postulates and requires for the main part only one variable, V_i. To this extent, then, model G is preferable. The testing of more detailed predictions will have to wait upon the collection of more precise data.

The nature of the model underlying theory G is such that it is impossible to talk about the results being due either to ' stimulus effects ' or ' response effects '. Because of what are traditionally called ' response predispositions ', the system requires less stimulus information to produce the same internal event for a high-frequency word than for a low-frequency word. It is not necessary, however, to talk in terms of a two-stage process; the system described in theory G is a single-stage process, as far as the word-frequency effect is concerned. Equally, the discussion as to whether the word-frequency effect has its origins in the frequency of experience or the frequency of emission of the words (Rubenstein & Aborn, 1960) is unnecessary. Within the preferred system, such a distinction becomes irrelevant (Morton, 1964; Morton & Broadbent, 1967). These conclusions agree with Broadbent (1967), who examined theories A, C, F and G, concluding that only theory G could account for his data.

The model underlying theory G can also be applied to the interaction of stimulus and context information in word recognition and the predictions derived from the theory are reasonably upheld (Morton, 1967). One version of this model has been stated axiomatically by Morton (1969).

APPENDIX

In discussing their Principle of Acoustic Equivalence, Brown & Rubenstein claim ' a response from a particular interval will be given just as often to words from one (different) interval as to words from another (different) interval '. They express this statement as the equation:

$$e_{i/j} = br_i + d \quad (j \neq i),$$

where $e_{i/j}$ is the probability of a response in the ith interval to a stimulus in the jth interval.

In fact, the above equation does not follow from the principle of acoustical equivalence. Consider the case of four word-frequency intervals which gives rise to the matrix in Table 4, where the entries correspond to the number of responses; and w, x, y and z are responses in the correct interval. Brown & Rubenstein's equation implies that $a=h=k$, and $b=e=l$, since $e_{i/j}$ is constant for j, the (incorrect) stimulus interval. The principle of acoustical equivalence, however, only implies that the frequency interval of an unrecognized stimulus has no effect upon the frequency interval of the response. That is, that the *ratios* of responses in two incorrect intervals remain constant; thus $a/b=k/l$. The equations they subsequently derive and their test of their data are thus invalid.

TABLE 4

		Response interval				
		1	2	3	4	Total
Stimulus interval	1	*w*	*a*	*b*	*c*	*N*
	2	*d*	*x*	*e*	*f*	*N*
	3	*g*	*h*	*y*	*i*	*N*
	4	*j*	*k*	*l*	*z*	*N*

ACKNOWLEDGEMENTS

This article was partly written while the author was on leave at the Center for Research on Language and Language Behavior, Ann Arbor, Michigan, U.S.A. and supported by a grant from the U.S. Office of Education. I am grateful to Dr H. Rubenstein for making his original data available to me. This article has benefited from discussions with Dr D. E. Broadbent.

REFERENCES

BROADBENT, D. E. (1967). Word-frequency effect and response bias. *Psychol. Rev.* **74,** 1–15.

BROWN, C. R. & RUBENSTEIN, H. (1961). Test of response bias explanation of word-frequency effect. *Science N.Y.* **133,** 280–281.

GREEN, D. M. & SWETS, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

HALLE, M. & STEVENS, K. N. (1962). Speech recognition: A model and a program for research. *I.R.E. Trans. Inf. Theory* **IT-8,** 155–159.

LIBERMAN, A. M., COOPER, F. S., HARRIS, K. S. & MACNEILAGE, P. F. (1963). A motor theory of speech perception. In C. G. M. Fant (ed.), *Proceedings of the Speech Communication Seminar.* Stockholm: Speech Transmission Laboratories, Royal Institute of Technology.

LUCE, R. D. (1959). *Individual Choice Behavior.* New York: Wiley.

MORTON, J. (1964). A preliminary functional model for language behaviour. *Int. Audiol.* **3,** 216–225.

MORTON, J. (1969). On the interaction of information in word recognition. *Psychol. Rev.* **76,** (in press).

MORTON, J. (1968). Grammar and computation in language behaviour. In *Progress Report No.* 5, C.R.L.L.B., University of Michigan.

MORTON, J. & BROADBENT, D. E. (1967). Passive vs. active recognition models or Is your homunculus really necessary? In W. Wathen-Dunn (ed.), *Models for the Perception of Speech and Visual Form.* Cambridge, Mass.: M.I.T. Press.

RUBENSTEIN, H. & ABORN, M. (1960). Psycholinguistics. *Ann. Rev. Psychol.* **11,** 291–322.

Vol. 21
Part 1
35–60

The British Journal of Mathematical and
Statistical Psychology

May
1968

COST, DISCRIMINABILITY AND RESPONSE BIAS

By PHILIP TWITCHELL SMITH

Institute of Experimental Psychology, Oxford and The Queen's College, Oxford

Cost, a unidimensional additive non-negative parameter, is defined and developed. Four models for the cost incurred in making a psychophysical decision are presented; two of these models are based on signal detectability theory and two are based on choice theory. In all models, cost is a function of the discriminability of the stimuli presented and the observer's response bias, and it is shown that in three of the models the form of this function depends critically on discriminability. The models are applied to the analysis of a reaction-time study, a vigilance study, and the statistical properties of the consonant systems of four natural languages.

1. INTRODUCTION

The purpose of this paper is to develop the concept of cost, and to apply this concept to the analysis of skilled human performance. The motivation for this approach arises from the belief that psychologists have not been sufficiently thorough in their application of terms like cost, effort, capacity and efficiency to the analysis of complex human performance. This point will be developed, with particular reference to psycholinguistics, in later papers.

The word ' cost ' is used to indicate a universal currency in terms of which all of a subject's actions (or at least an important subset of them) can be evaluated: it rests on the intuitively attractive assumption that there is a unidimensional parameter that can be attached to every action and which is fundamental in determining the bounds of further courses of action. This concept is closely related to what some psychologists (James, 1890; Thurstone, 1930; Zipf, 1949) have called effort; it is a generalization of the notion of pay-off in signal detectability theory and utility theory, and it is used in the same spirit as some psychologists (e.g. Garner, 1962) have used an information measure.

The basic notion behind the cost concept is that when a human subject makes a decision or carries out an action, he does so in terms of a sequence of fundamental processes. A cost (real-valued parameter) can be attached to each fundamental process independent of all other processes, and the total cost of a sequence of fundamental processes is the *sum* of the costs for each individual fundamental process: at the level of fundamental processes, cost is additive. A second requirement for a cost function is that costs derived from quite different systems of fundamental processes shall be commensurable, i.e. cost is unidimensional.

As an example, consider a desk calculating machine. Although the machine can be pre-programmed to carry out addition, subtraction, multiplication, division and other more complicated operations, usually such a machine is basically only a device that can carry out repeated addition operations in rapid succession, together with a number of procedures for converting multiplication, division, etc. into a series of addition procedures. Thus the fundamental processes of this machine would be the addition procedure and the transitions between successive addition procedures. (There might be several forms of transition that should be dealt with separately.) The cost of a complete calculation could then be defined as the sum of the costs for each addition procedure plus the costs for each transition. Identifying the fundamental processes of the system still leaves open the precise measurement of cost: in the present example, the time it takes to complete the calculation or the wear and tear to the hardware of the machine would be ' natural ' ways of defining cost if exclusive interest was centred on speed or durability respectively, and both these measures would satisfy the requirement of additivity under the identification of the fundamental processes. However, conversely, the choice of a particular cost measure might radically alter the identification of the fundamental processes: in a machine that can carry out a variable number of operations in parallel, the fundamental processes that guarantee an additive measure of cost in terms of wear and tear will not necessarily produce an additive measure of cost in terms of time (n operations in parallel could take the same time no matter what value n takes, whereas wear and tear could be proportional to n).

It would be reasonable to expect a person wishing to have a calculation performed for him to pay for all the sources of cost in the task: if it requires a skilled operator to operate the machine, which is liable to wear and tear, then the customer should pay both the operator's wages (which might be worked out on a time basis) and for the replacement of worn-out parts of the machine. The fact that both time and wear and tear can be reduced to a single quantity in this case—money—is what is meant by the requirement that costs from all sources shall be commensurable.

2. Models of Cost

Many previous models that have been implicitly concerned with cost have been ' repeated sample ' models: information is collected by taking repeated independent samples of the stimulus presented, until the receiver has enough relevant information to make an appropriate response. In these circumstances, the process of taking a single sample can be regarded as a fundamental process and cost is closely related to sample size. Models of this kind are, for example, the zero-memory model of Swets & Green (1961), the repeated covert response model of Audley (1960), and the sequential testing model of Stone (1960). This paper, however, is concerned with ' single sample ' models, where the fundamental processes of the cost system are steps in a calculation based on a single item of information, or where cost reflects some internal state of the receiver. These models will be based on two psychophysical theories: signal detectability theory and choice theory. A full discussion of these two theories can be found in Luce (1963).

2.1. *Models based on Signal Detectability Theory*

In signal detectability theory it is assumed that each stimulus presented to a subject can be assigned a value by the subject along a unidimensional continuum, which will be called the *decision axis*. The subject knows *a priori* (probably by previous experience) the probability distributions of these values for any given stimulus. (In general, different presentations of the same stimulus produce different values, because of imperfections in the sensory system.) Both theory and experiment suggest that these probability distributions are normal distributions. When the subject has to judge which of two possible stimuli has been presented to him, he is assumed to use a criterion level on the decision axis, such that if the value for the stimulus presented exceeds this level, one stimulus is reported, and if the value is less than this level, the other stimulus is reported.

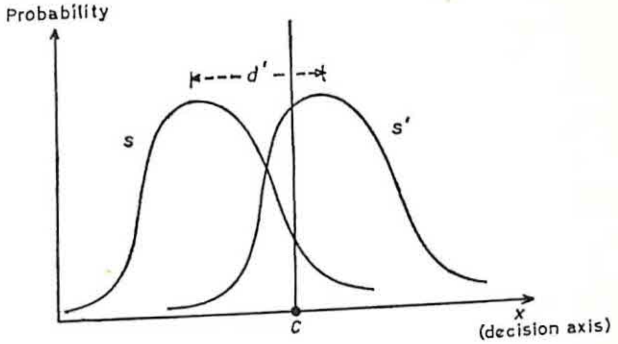


FIGURE 1. Probability distribution of two stimuli along a unidimensional decision axis.

Thus in Fig. 1, the curves labelled s and s' represent the probability distribution of values on the decision axis (x axis) for the stimuli s and s'. The subject will respond to the stimulus s' if the value of the stimulus presented exceeds C on the decision axis, and to stimulus s if the value of the stimulus presented is less than C. The particular position of C is determined by pay-off conditions (the rewards and punishments for the various correct and incorrect responses).

For the models of cost in this context, it will be assumed that collecting information about the probability distribution does not contribute to the cost incurred by a subject. (In the computer example this is preliminary programming, which has already been carried out before the task begins.) In this section the system arising when the decision is made from a single sample of the value of the stimulus presented will be considered.

There are two possibilities: that making an individual decision costs the same no matter what the situation is, or else that some decisions are inherently more difficult and costly than others. This second possibility could arise as follows: suppose that determining the value of the x co-ordinate for a particular stimulus presentation involves cost, and the more precise the value calculated the more the cost. (This situation is completely analogous to carrying out a long-division sum by hand—it requires much more effort to determine the answer to 10 significant figures than it does to only two significant figures.) If this x co-ordinate is considerably above or below the criterion level, this will become apparent after only a small amount of calculation, and a decision can be made; however, if the x co-ordinate is close to the criterion level, then extra calculation will be needed to determine more precisely whether it is just above or just below this level. Thus, in this model, the criterion level is at a precise location, and the x co-ordinate, the value associated with the particular stimulus presentation, is ill-defined and it involves cost to make it more precise. There is also a complementary version: the x co-ordinate is precise but the criterion level is vague and is only specified more precisely when a close decision has to be made. Whichever way one looks at the problem, the same conclusion appears: cost is a monotonic decreasing function of distance from the criterion level, which is independent of the distributions of the stimulus values being considered.

Two precise versions of the notion will be discussed. The first follows the analogy with the determination of significant figures in a calculation and gives cost as a logarithmic function of distance from the criterion level; the second, which is mathematically easier to handle, is a two-level step function.

Model 1 Suppose values on the decision axis are expressed as a-ary numbers, which almost never exceed the value a^n for some fixed n, i.e. values can be represented as numbers of the form $k \times a^{n-1}$, where $0 \leqslant k < a$. Suppose the criterion level value is C, and a particular stimulus presentation yields a value differing from C by an amount ρ. Now ρ can be represented as an a-ary number of the form $k' \times a \,\mathrm{int}(\log_a \rho)$ (where $1 \leqslant k' < a$ and $\mathrm{int}(\log_a \rho)$ is the greatest integer less than $\log_a \rho$). Therefore, if the value on the decision axis is calculated by first determining the coefficient of a^{n-1}, then the coefficient of a^{n-2}, etc., we shall need to make $n - \mathrm{int}(\log_a \rho)$ such determinations before we can be certain that the value of the stimulus presented is above or below the criterion level. (For example, in the decimal system, suppose no value on our scale exceeded 10,000, i.e. $n=4$, the criterion level was 6,235, and the stimulus value was 6,256, i.e. $\rho=21$ and $\mathrm{int}(\log_{10} 21)=1$; then it is necessary to determine the stimulus value to three significant figures before it is known that $6{,}256 - 6{,}235$ is a positive quantity.)

If it is postulated that cost is proportional to the number of significant figures calculated, i.e. each calculation of each significant figure is a fundamental process with the same cost, then by choosing appropriate units:

$$\text{cost} = E \text{ (say)} = n - \mathrm{int}(\log_a \rho).$$

Since the choice of a is arbitrary, the above expression can be replaced by the continuous approximation $E = K - \log \rho$, where it is understood that $\log \rho < K$, always.

Model 2 This is simpler, but does not make explicit the underlying fundamental processes. It is assumed that there is a certain critical region around the criterion level where extra cost is involved. Outside the critical area the cost is minimal, i.e.

$$E = E_0 \quad (\rho > k),$$
$$E = E_1 \quad (\rho \leqslant k;\ E_1 > E_0).$$

The curves of cost against stimulus value are shown for the two models in Fig. 2.

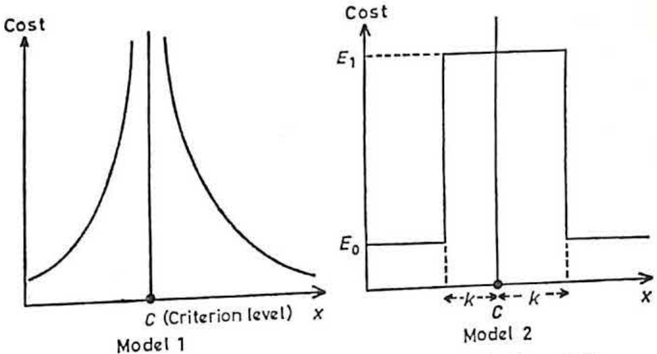


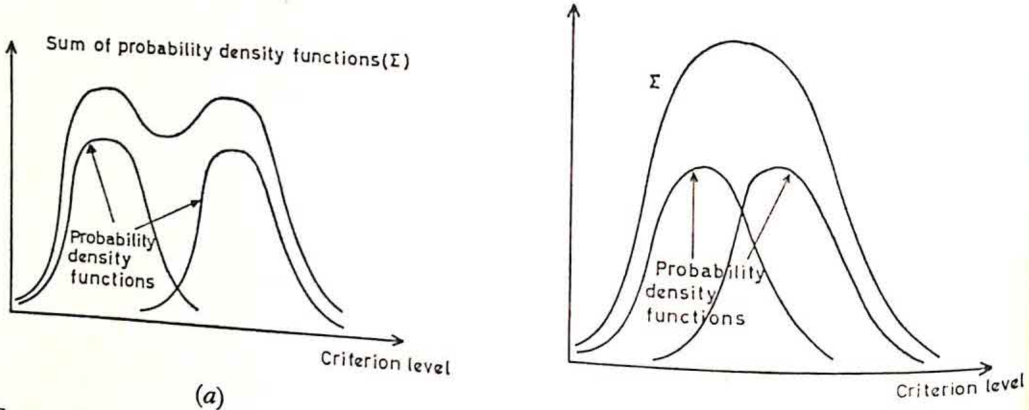
FIGURE 2. Cost functions for models 1 and 2.

Neither model is easy to handle if exact solutions are required, but since both models give approximately the same results, and since approximate solutions, as will be seen, are most easily handled by model 2, attention will be confined to this.

Some results about *average* cost in a given task can be derived. For model 2, the average cost is a linear function of the proportion of decisions the subject makes that are in the critical region around the criterion level. If this critical region is small compared to the standard deviations of the stimulus value probability distributions, then to a good approximation the proportion of critical decisions = width of critical region $(2k) \times$ the sum of the probability density functions at $C(\Sigma)$, i.e. $E = E_0(1 - 2k\Sigma) + 2kE_1\Sigma$. If, for a given sensory system, it is assumed that k is constant, this gives the result that *average cost is a linear function of the sum of the probability density functions at the criterion level.*

This leads to some interesting results. Consider two signals, s, s', each of whose stimulus value distributions has variance σ^2. If the means of these distributions are separated by a distance d and either signal is equally likely to occur, this gives the situation pictured in Fig. 1, with $d' = d/\sigma$. If d' is sufficiently large, the average cost as the criterion level is changed will be less for central values of criterion level in between the peaks of the two distributions, since the sum of the probability density functions is smaller here than at one of the peaks, and thus less critical decisions on average will be made. If d' is small, however, cost will be larger for central values of criterion. This is illustrated by Fig. 3.

The dip in the cost curve shown in Fig. 3 (*a*) appears for values of d' greater than 2 (for proof, see Appendix 1). This result has the consequence that for easy discriminations (large d') it involves less cost to be unbiased, but the reverse is true for difficult discriminations (small d'). Table 1 lists the sums of probability density functions for two equiprobable signals, for various criterion levels and values of d'.


Sum of probability density functions(Σ)
Probability density functions
Criterion level
Σ
Probability density functions
Criterion level

(*a*) (*b*)

FIGURE 3. Sum of probability density functions for two stimuli: (*a*) separated by large d'; (*b*) separated by small d'.

TABLE 1. SUMS OF PROBABILITY DENSITY FUNCTIONS OF TWO EQUIPROBABLE STIMULI

(β is the value of the probability ratio at the criterion point C.)

d'	β 1	1·5	2	3	4
0	0·80*				
0·5	0·77	0·55	0·30	0·07	0·02
1	0·70	0·66	0·59	0·44	0·33
1·5	0·60	0·59	0·57	0·53	0·48
2	0·48	0·48	0·48	0·48	0·48
2·5	0·36	0·37	0·37	0·38	0·39
3	0·26	0·26	0·27	0·28	0·30

* Strictly speaking, this is the limit as $d' \to 0$ of Σ for $\beta = 1$.

2.2. *Models based on Choice Theory*

The second class of models to be discussed arises from the psychophysical model based on choice theory (Luce, 1963). For a given psychophysical task a stimulus set S, a response set R and an identification function ι need to be defined. ι is a mapping of S into R, and for $s \in S$ and $r \in R$ $\iota(s) = r$ if r is a correct response to s, where 'correctness' is defined by the experimenter at the beginning of the task. The most general model that can be developed from this formulation is derived by postulating a function A of S, R, s, r, ι, for each stimulus s and response r, such that the probability of making the response r given the stimulus s is

$$p(r|s) \propto A(S, R, s, r, \iota).$$

This model is made more specific by supposing

$$p(r|s) \propto A(s, r, \iota), \tag{1}$$

i.e. the probability of making the response r given the stimulus s is proportional to a number determined by s, r, and the identification function ι only: the alternative stimuli to s in S and the alternative responses to r in R are considered irrelevant to determining the size of $A(s, r, \iota)$.

The proportionality sign in eqn. (1) can be removed by rewriting this equation as

$$p(r|s) = \frac{A(s, r, \iota)}{\sum\limits_{r' \in R} A(s, r', \iota)}. \tag{2}$$

Eqn. (2) can be made yet more specific by assuming

$$A(s, r, \iota) = \eta(s, \iota^{-1}(r))b(r),$$

where η is a mapping of $S \times S$ into the non-negative numbers, and b is a mapping of R into the non-negative numbers. η is a similarity function: the more similar s and s' are, the larger $\eta(s, s')$ is; and b is a response bias function: the greater the bias towards a particular response r, the greater is $b(r)$. Further assumptions are usually made: the most common is that $-\log \eta$ is a metric on $S \times S$.

This is not the place for a critical review of this model: for a more careful outline of the assumptions in this model, the reasons for making them and a review of experimental work relevant to the model, see Luce (1963).

For our purposes it need only be noted that (1) the model is probabilistic: there is no way of determining on a particular trial what the actual response will be with certainty; (2) it is a summary of performance, not a description of it: in the present terminology, the fundamental processes are not explicit; (3) the complicated notions of stimulus discriminability and response bias are each handled by single numerical quantities, and these quantities are combined together according to simple arithmetical rules. The models of cost to be considered will be concerned (1) with the cost in making choices given a set of probabilities, (2) with the cost associated with computing and maintaining bias parameters.

2.3. *Response Entropy Model*

Model 3 This deals with cost in a choice situation where the response probabilities are given, and simply postulates that cost is proportional to entropy. The system is as follows. Given an unidentified stimulus s, the receiver computes a set of probabilities $\{p(r|s)\}$ specifying the probability of each response r

(the present choice model predicts what these probabilities will be without specifying the mechanism by which they were obtained). It is assumed that this computation involves negligible or constant cost. The receiver's task is then to select a response r, subject to the restriction that the set of possible responses has the probability distribution just specified. For example, suppose $p(r_1|s)=p(r_2|s)=\frac{1}{2}$, and $p(r_i|s)=0$ for $i \geqslant 3$, i.e. if the receiver were required to make a long sequence of independent responses in this situation, 50 per cent of his responses would be r_1 and 50 per cent would be r_2: in this case the receiver would need, mentally, to toss a coin to decide between r_1 and r_2. The notion of coin-tossing can be formalized in the general case by supposing the receiver has some system of generating random digits. When a particular sequence of digits is generated, this produces a signal to emit a particular response. For example, consider four possible responses r_1, r_2, r_3, r_4, to a stimulus s, with $p(r_1|s)=\frac{1}{2}$, $p(r_2|s)=\frac{1}{4}$ and $p(r_3|s)=p(r_4|s)=\frac{1}{8}$, and a system for generating binary random digits. The receiver could decide to respond with r_1 after the sequence 0, r_2 after 10, r_3 after 110 or r_4 after 111, since these sequences of binary digits would occur with probabilities $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ and $\frac{1}{8}$ respectively. A method for constructing such sequences is essentially the same as the Huffman method of constructing a minimum redundancy code (Huffman, 1952; Abramson, 1963). This leads to the result that the average length of the binary sequence required by this decision rule is

$$-\sum_s\sum_r p(s)p(r|s)\log_2 p(r|s),$$

and more generally, the familiar theorem of Shannon (1948) allows us to say that for any probability distribution and sequence of a-ary random digits, the average length of digit sequence needed to reach a decision is

$$-\sum_s\sum_r p(s)p(r|s)\log_a p(r|s),$$

and moreover, the length of sequence when a particular response r' is made to a particular stimulus s' is $-\log_a p(r'|s')$.

On a descriptive level this handles many experimental results on reaction time quite well. Consider the extreme cases of perfect discriminability and zero discriminability. For perfect discriminability

$$\eta(s, s')=\delta(s, s'),$$

where δ (the Kronecker δ) is such that

$$\delta(x, y)=1 \quad \text{if} \quad x=y, \qquad =0 \quad \text{otherwise.}$$

Thus

$$p(r|s)=\delta(s, \iota^{-1}(r)) \quad \text{and} \quad \sum_r -p(r|s)\log p(r|s)=0.$$

This implies that cost (and hence reaction time) is at a minimum value that is the same for all responses and stimuli and is independent of the number of responses or the number and probability of the stimuli.

For zero discriminability $\eta(s, s')=k\neq 0$ independent of s and s', thus $p(r|s)=p(r)$ and

$$-\sum_s\sum_r p(s)p(r|s)\log p(r|s) = -\sum_r p(r)\log p(r).$$

In this case, cost and reaction time are linear functions of response entropy. Superficially at least, this accords well with experimental results (e.g. Garner, 1962, ch. 2).

It will be useful to compare the properties of this model with the corresponding signal detectability model in the previous section.

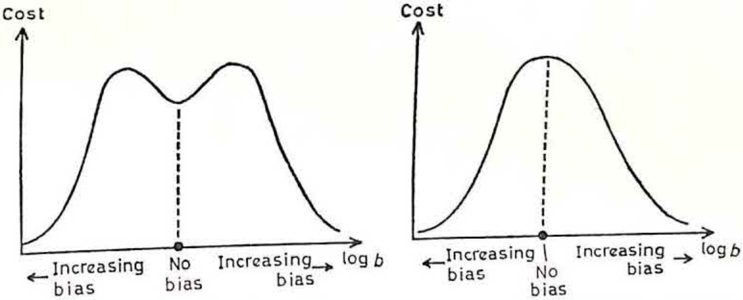
To make a direct comparison with the signal detectability model, consider two equiprobable stimuli s, s' to which the appropriate responses are r and r'. Taking $\eta(s, s)=\eta(s', s')$, and writing

$$\frac{\eta(s, s)}{\eta(s, s')}=\eta, \qquad \frac{b(r)}{b(r')}=b,$$

it is not difficult to derive (see Appendix 2) that

$$\text{mean cost} = \bar{E} = \tfrac{1}{2}\left\{\log(\eta b+1)+\log(\eta+b) - \eta\left[\frac{1+2\eta b+b^2}{(\eta b+1)\,(\eta+b)}\right]\log\eta - b\left[\frac{1+2\eta b+\eta^2}{(\eta b+1)\,(\eta+b)}\right]\log b\right\}$$

and $\bar{E}=0$ when $b=0$ or ∞ (the cases of maximum differential bias between responses). Moreover, when $b=1$ (the case of no differential bias between responses) $\bar{E}$ is a maximum if $\eta<4{\cdot}681$ and a minimum if $\eta>4{\cdot}681$ (Appendix 2). This leads to dromedary-shaped and camel-shaped curves similar to those obtained for models 1 and 2 (see Fig. 4).



High discriminability $\eta>4{\cdot}681$ Low discriminability $\eta<4{\cdot}681$

FIGURE 4. Cost as a function of response bias for model 3.

Note that the critical points (for transition between the bimodal and unimodal distributions) for the signal detectability models ($d'=2$) and for the choice theory model ($\eta=4{\cdot}681$) are quite closely equivalent. For example, the minimum error rate at the critical point on the signal detectability models is 15·8 per cent and on the choice theory model it is 17·6 per cent.

Not only are the shapes of the cost–bias curves much the same for the two models, but the variation of cost with discriminability is comparable for the models. Table 2 compares cost predicted by the two models for two equiprobable stimuli when there is no differential bias between responses, for various error rates: the range of cost predicted is slightly smaller for the choice theory model, but the two models give very similar results for low error rates. It is well known (e.g. Luce, 1963) that the ROC curves predicted by signal detectability theory and choice theory are also very similar, and one way of putting this is that, although based on very different concepts, the two models are in some sense equivalent. The results on cost that have been derived further suggest that distance from criterion level in signal detectability terms is equivalent to logarithm of probability in choice theory terms.

TABLE 2. COST PREDICTED BY MODELS 2 AND 3 FOR DIFFERENT MINIMUM ERROR RATES

(Cost is taken, with error rate at the 'critical points' of each model, as unity; i.e. at $d'=2$ for model 2 and at $\eta=4{\cdot}681$ for model 3. The error rates are those obtained for multiples of $\frac{1}{2}d'$.)

Error rate	Cost	
	Signal detectability theory (model 2)	Choice theory (model 3)
0·500	1·67	1·48
0·401	1·60	1·44
0·309	1·46	1·32
0·227	1·25	1·14
0·159	1·00	0·94
0·106	0·75	0·72
0·067	0·54	0·53

2.4. *Cost and Bias Parameters*

At this point a new possible source of cost will be introduced. While the considerations to be outlined apply equally well to signal detectability theory models, they are more easily expressed in terms of choice theory.

The idea is as follows. There is a 'natural' level for each bias parameter: to maintain a bias parameter at any other level involves cost. For example, consider a 'Yes–No' choice (signal vs. noise) psychophysical task where the subject receives a small monetary reward for correct responses, but is heavily fined for failing to detect a signal. The strategy that maximizes the subject's expectation of reward from a long series of trials is one where an extreme criterion level is adopted—the subject says 'Yes' almost all the time, and says 'No' only when he is exceptionally certain that no signal was presented. The suggestion is that in this case the subject incurs cost in accepting stimuli that he would 'normally' reject as signals. Some experimental evidence can be taken as support for this view: Norman (1966, personal communication) reports that although subjects in such situations are aware of the necessity for maintaining extreme criterion levels, there is a tendency over a series of trials for them to recede to a more moderate criterion.

The fact that choice theory is a summary and not a description of performance again means that the fundamental processes are not obvious and only the form of the cost function will be proposed. It is postulated that $B(r)$ is the 'natural' level of $b(r)$ and that for $b(r) > B(r)$, cost is a monotonic increasing function of $b(r) - B(r)$, and for $b(r) < B(r)$, cost increases monotonically with $B(r) - b(r)$. It is difficult to be more specific than this without making dubious assumptions.

Model 4 For explicitness this model makes a few such assumptions.

(1) In the two-choice case, if $b(r) = B(r)$ and $b(r') = B(r')$, the cost involved in increasing $b(r)$ to $kb(r)$ is the same as that involved in decreasing $b(r')$ to $k^{-1}b(r')$. (Note that both these changes have the same effect on the set of probabilities $\{p(r|s)\}$.) This suggests that $\log b(r)$ rather than $b(r)$ is a more convenient unit with which to measure cost.

(2) Cost associated with maintaining $b(r)$ at a level different from its natural level $B(r)$ is

$$E = \{\log [b(r)/B(r)]\}^2.$$

This is a simple mathematically convenient function that satisfies all the considerations outlined so far.

In the two-choice case, cost E can be taken as

$$E = [\log (b(r)/B(r))]^2 + [\log (b(r')/B(r'))]^2.$$

The usual minimization procedures show that for $B(r)$ and $B(r')$ fixed and $b(r)/b(r') = b$ kept constant, E is a minimum when

$$\log b(r) = \tfrac{1}{2} \log b + \tfrac{1}{2}[\log B(r) + \log B(r')],$$
$$\log b(r') = -\tfrac{1}{2} \log b + \tfrac{1}{2}[\log B(r) + \log B(r')],$$

i.e. the logarithms of the bias parameters are $+\frac{1}{2} \log b$ and $-\frac{1}{2} \log b$ away from the mean of the logarithms of the 'natural' levels of the bias parameters. In this case

$$\min E = \tfrac{1}{2}\{[\log b - \log B(r) + \log B(r')]^2\}.$$

Fig. 5 shows the variation of $\min E$ with $\log b$. Cost is zero when $b(r) = B(r)$, $b(r') = B(r')$ and $\log b = \log B(r) - \log B(r')$. This model can be applied to the problem of distinguishing a pair of stimuli of unequal frequency. In this case, if $p(s) = P$ and $p(s') = 1 - P$,

$$\text{error rate} = Pp(r'|s) + (1 - P)p(r|s')$$
$$= P\frac{1}{\eta b + 1} + (1 - P)\frac{b}{\eta + b}$$

(from the results of Appendix 2).

It can be shown (e.g. Luce, 1963) that optimum b is then given by

$$b=\frac{\eta-\sqrt{[(1-P)/P]}}{\eta\sqrt{[(1-P)/P]}-1}, \tag{3}$$

provided this leads to a result giving $0<b<\infty$; otherwise $b=0$ or ∞.

Now min $E\sim\frac{1}{2}(\log b)^2$ for b large or small compared with unit, and for η fixed and sufficiently large or small P, optimum b, as given by eqn. (3), can be made arbitrarily large or small. This means that if in practice there is a limit to the total cost that can be incurred, optimum values of b are simply unattainable for extreme values of P. In other words, for fixed η, efficiency, as measured by how far the error rate departs from the minimum obtainable error rate, will be less in the two-choice case when the probabilities of the two stimuli are very different. Experimental data supporting this statement are provided by Green (1960).

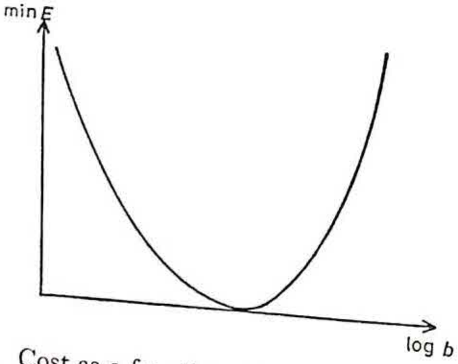


FIGURE 5. Cost as a function of response bias for model 4.

Another convenient feature of model 4 is that it is formulated in such a way as to handle sequential effects quite neatly. It could be assumed that the 'natural' level of bias $B(r)$ increases shortly after the response r has been made, returning to its original level gradually over time: this sort of model would handle the effects of 'preparation' and 'expectancy' encountered in reaction time studies (e.g. Bertelson, 1961).

3. THE RELATION BETWEEN COST AND PAY-OFF

Besides the requirement of additivity, the other critical feature of cost is that costs derived from different sources shall be commensurable. In the discussion so far, the description of the processes of decision has been simplified in such a way that the cost of a decision can be expressed as a not too complicated function of discriminability, response bias and error rate: in such a preliminary investigation, the natural additional simplification to make for studying commensurability is that all the *consequences* of a decision can be reflected in a single pay-off matrix whose elements are commensurable with the cost of making the decision. The cost of making a decision of a certain accuracy can then be balanced against the reward to be received for such accuracy—in certain cases it is conceivable that the reward for accuracy is so small that it is not worth entering a costly decision process to improve accuracy or, at the other extreme, it is possible that accuracy is so important that it is never worth terminating a decision procedure while greater accuracy can still be attained.

3.1. *Cost of Decision vs. Improved Accuracy of Response for Models* 1 *and* 2

Consider a zone of uncertainty of length h, in which the criterion level lies, where the observer's calculating limitations do not allow him to determine, when the likelihood ratio falls in this region, whether it is above or below criterion. Outside this region, accurate determination is always possible.

If c is criterion, and $f_Y(x)$ and $f_N(x)$ are the probability density functions for signal and noise respectively, then if there were no calculating limitations,

$$p(Y|s)=\int_c^\infty f_Y(x)dx. \tag{4}$$

If $p(Y'|s)$ is the probability of responding ' Yes ' to signal when such calculating difficulties exist, and $p(Y'|s, g)$ is the probability of responding ' Yes ' to signal in the particular case when the criterion level is distance g from the lower bound of the zone of uncertainty, then if it is assumed that, over repeated trials, the criterion level is randomly and uniformly distributed within the zone of uncertainty,

$$p(Y|s)=\frac{1}{h}\int_0^h p(Y'|s, g)\,dg \tag{5}$$

and

$$p(Y'|s, g)=\int_{c+h-g}^\infty f_Y(x)\,dx+\alpha\int_{c-g}^{c+h-g} f_Y(x)\,dx, \tag{6}$$

where α is the proportion of ' Yes ' responses made in the zone of uncertainty. If it is assumed that the observer has no information about the position of the zone of uncertainty relative to the criterion level, then α will be independent of g. Writing

$$F_Y(x')=\int_{-\infty}^{x'} f_Y(x)\,dx$$

and observing that

$$F_Y(c+z)=F_Y(c)+zf_Y(c)+\tfrac{1}{2}z^2f_Y'(c)$$

to second order in z, by Taylor's theorem, then eqn. (6) can be rewritten as:

$$\begin{aligned}P(Y'|s, g)&=1-F_Y(c+h-g)+\alpha[F_Y(c+h-g)+F_Y(c-g)]\\&=1-F_Y(c)-(1-\alpha)[(h-g)f_Y(c)+\tfrac{1}{2}(h-g)^2f_Y'(c)]\\&\quad+\alpha[gf_Y(c)-\tfrac{1}{2}g^2f_Y'(c)]\\&=p(Y|s)-[h(1-\alpha)+g]f_Y(c)-\tfrac{1}{2}[h^2(1-\alpha)-2hg(1-\alpha)+g^2]f_Y'(c).\end{aligned}$$

Applying eqn. (5) gives:

$$p(Y'|s)=p(Y|s)+\tfrac{1}{2}h(2\alpha-1)f_Y(c)-h^2f_Y'(c)/6.$$

Similarly,

$$p(Y'|n)=p(Y|n)+\tfrac{1}{2}h(2\alpha-1)f_N(c)-h^2f_Y'(c)/6.$$

Suppose the pay-off matrix is as follows:

$$\begin{array}{ccc} & Y & N \\ s & o_{11} & o_{12} \\ n & o_{21} & o_{22} \end{array}$$

Then the expected gain in reward (G) when $p(Y'|s)$, $p(Y'|n)$ are used instead of $p(Y|s)$, $p(Y|n)$ is:

$$\begin{aligned} G=\tfrac{1}{2}h(2\alpha-1)f_Y(c)P(o_{11}-o_{12})-h^2f_Y'(c)P(o_{11}-o_{12})/6 \\ +\tfrac{1}{2}h(2\alpha-1)f_N(c)(1-P)(o_{21}-o_{22}) \\ -h^2f_N'(c)(1-P)(o_{21}-o_{22})/6. \end{aligned} \tag{7}$$

Now if criterion is optimum, i.e. if it maximizes expected reward,

$$\frac{f_Y(c)}{f_N(c)}=\frac{1-P}{P}\,\frac{o_{22}-o_{21}}{o_{11}-o_{12}}.$$

(This is a standard result of signal detectability theory, e.g. see Luce, 1963, pp. 148–149.) Therefore, terms in h cancel out in eqn. (7), giving the result:

$$G=-h^2(Rf_Y'(c)-Sf_N'(c)), \tag{8}$$

where R, S are positive constants determined by the *a priori* probabilities and the pay-off matrix. Note that in the symmetrical unbiased case, where $R=S$,

$$\begin{aligned} f_Y(x)&=(2\pi)^{-1/2}\exp\left(-\tfrac{1}{2}(x-\tfrac{1}{2}d')^2\right), \\ f_N(x)&=(2\pi)^{-1/2}\exp\left(-\tfrac{1}{2}(x+\tfrac{1}{2}d')^2\right) \quad \text{and} \quad c=0, \\ Rf_Y'(c)-Sf_N'(c)&=(2\pi)^{-1/2}Rd'\exp\left(-d'^2/8\right), \end{aligned}$$

which is a maximum at $d'=2$ (by simple calculus) and $\to 0$ as $d'\to 0$ or ∞. That is, the effect of inaccuracy is negligible for poor discriminability (things cannot get worse) and for high discriminability (things are too easy), and maximum effect occurs for intermediate discriminability ($d'=2$).

The conclusions of this section are that an observer who tries to balance the cost of making a decision of a certain accuracy against the consequences of such a decision again finds $d'=2$ a critical point. Whereas in the previous section $d'=2$ was the transition point between unimodal and bimodal cost–response bias curves, in this section $d'=2$ is the point where inaccuracy has its largest effect for equiprobable stimuli and an unbiased pay-off matrix.

4. Some Applications

There are difficulties in attempting to apply some of the results that have been derived for commensurability: the measurement of h, the width of the zone of uncertainty, would involve some cumbersome techniques: in a confidence rating experiment it would mean asking the subject how confident he was about his confidence ratings, and in a reaction time experiment, the supposition that the subject goes on calculating until he reaches the limit of accuracy determined by h, would imply that h could be inferred from the upper limit of reaction times, when artifacts such as the subject falling asleep or otherwise failing to respond are allowed for: obviously this sort of procedure would seem arbitrary. For these reasons, no attempt is made in this paper to review experimentally the hypothesized relation between the accuracy parameter h of models 1 and 2 and discriminability. Some discussion of the commensurability aspects of cost will be attempted in a later paper on communication with artificial languages.

In the examples that follow, the discussion is restricted to the demonstration that one of the cost models (model 2) can make relevant predictions in a number of fields.

(1) *Reaction time* The study of Carterette *et al.* (1965) relates, *inter alia*, discriminability and mean reaction time for different probabilities of a signal in noise. Taking mean reaction time as a measure of cost it is possible, at least qualitatively, to account for certain of their results quite neatly.

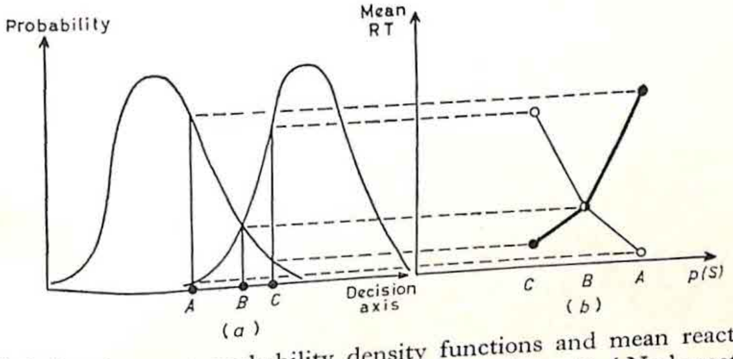


Figure 6. Relation between probability density functions and mean reaction time for stimuli of high discriminability. ○, ' Yes ' reaction time; ●, ' No ' reaction time.

Using model 2, since it can most readily be represented graphically, Figs. 6 (*a*) and 7 (*a*) show the cost associated with a lax criterion A (high probability of saying ' Yes '), an unbiased criterion B, and a cautious criterion C (low probability of saying ' Yes '). It is assumed that the same three values of β (the likelihood ratio) are used to determine the positions of A, B, C in both figures. These positions have been chosen *a posteriori* to *fit* the data rather than *a priori* to *predict* the data. Assuming such different criteria are adopted as the *a priori* signal probability changes, and mean cost and mean reaction time are equivalent, Figs. 6 (*b*) and 7 (*b*) indicate the predicted variation of reaction time with *a priori* probability. The essential point is that for low discriminability (Fig. 7) the cost and reaction time for 'Yes' is higher for the unbiased criterion (*B*) than for either of the biased criteria (*A* or *C*); whereas in the high discriminability case (Fig. 6) the unbiased criterion gives a reaction time intermediate between the reaction times for the other criteria. Carterette *et al.* (1965) report data for three subjects: for two of them the mean reaction time – *a priori* signal probability curves are as in Fig. 6 (*b*), but for the other subject the curves are as in Fig. 7 (*b*): this subject has also markedly lower discrimination. This is mild support for model 2.

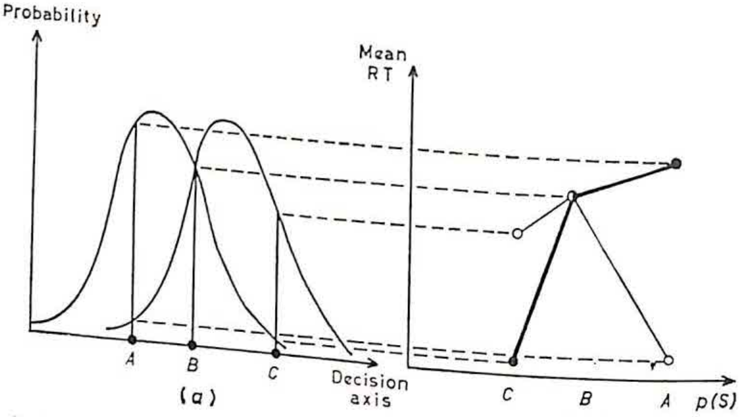


FIGURE 7. Relation between probability density functions and mean reaction time for stimuli of low discriminability. ○, 'Yes' reaction time; ●, 'No' reaction time.

(2) *Vigilance* Broadbent & Gregory (1963) analyse a vigilance task in terms of position of criterion level. Since the signal was infrequent, the situation can be represented by Fig. 8: consider a set of criteria *A*, *B*, *C*, *D*, *E*, ranging from lax (*A*) to cautious (*E*). Suppose the effect of fatigue in this task is for the subject to try to reduce cost by altering the criteria. He can reduce cost for his more cautious criteria, *C*, *D*, *E*, by making them more cautious, i.e. following the downward slope of the cost curve. He cannot do the same for his lax criterion (*A*), since to change its position by small amounts inevitably increases cost. It is therefore predicted that, in a prolonged vigilance task, cautious criteria will become more cautious, but lax criteria will remain the same. This is Broadbent & Gregory's findings. A less convincing prediction can be offered for the effects of the external distracting noise on the task. To resist noise, a subject must increase cost; (the attention mechanism works like an ideal economist and directs attention to those parts of the system where most cost is being incurred); increasing cost can be achieved in Fig. 8 (*b*) by making the criteria converge on *X*, i.e. by making the criteria closer together—this again is Broadbent & Gregory's finding.

(3) *Linguistics* As a final example, the relation between response bias and discriminability in the consonant systems of natural languages will be discussed.

Many linguists (e.g. Jakobson, 1953; Chomsky, 1965; Halle, 1958) have proposed that each phoneme should be represented as a collection of binary *distinctive features*. Results for the consonant system are presented which are based on one such feature analysis (Halle, 1958). While many linguists (including apparently Halle himself) would not consider this particular analysis the best in all its details (e.g. Halle, 1965, now uses in addition the feature *gradual release* for describing the consonants of English), the following argument is unlikely to be affected by minor adjustments in the details of this system. The Halle (1958) system is shown in Table 3 (with two extra consonants added). Response bias and discriminability will be looked at in this system for the following six cases.

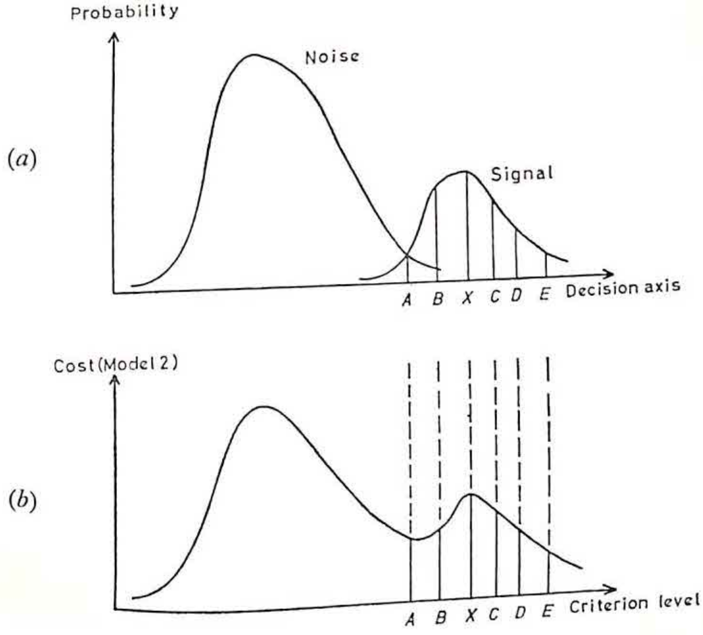


FIGURE 8. Cost derived from probability density functions in a vigilance task.

1. English consonants at the beginning of words and immediately preceding long *a* (*a* as in *father*). Source: all words occurring in the Thorndike–Lorge (1944) word count.

2. English consonants at the end of words and immediately following long *a*. Source: Thorndike–Lorge.

3. English consonants at the beginning of words and immediately preceding the liquids (*l* or *r*). Source: Thorndike–Lorge.

4. French consonants at the beginning of words and immediately preceding *a* (the vowel sound in the French words *chat* and *rare*). Source: *Harrap's Shorter French and English Dictionary* (1940).

5. German consonants at the beginning of words and immediately preceding *a* (the vowel sound in the German words *lamm and lahm*). Source: *Cassell's Compact German Dictionary* (15th edition, 1958).

6. Welsh consonants at the beginning of words (radical forms only, neglecting the mutations that occur in the initial consonants of some words). Source: *The New Welsh Dictionary* (1953).

TABLE 3. HALLE'S ANALYSIS OF THE 20 CONSONANTS* WHICH APPEAR IN THE PRESENT SAMPLES

p as in *p*ar
b as in *b*ar
m as in *m*ar
f as in *f*ar
v as in *v*ase
k as in *c*ar
g as in *g*arden
t as in *t*ar
d as in *d*ark
θ as in *th*igh
ð as in *th*y
n as in *n*ame
s as in *s*at
z as in *z*ebra
č as in *ch*ar
ǰ as in *j*ar
š as in *sh*arp
ž as in plea*s*ure
ŋ as in si*ng*
x as in Ba*ch*

	p	b	m	f	v	k	g	t	d	θ	ð	n	s	z	č	ǰ	š	ž	ŋ	x
grave	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	+	+
diffuse	+	+	+	+	+	−	−	+	+	+	+	+	+	+	−	−	−	−	−	−
strident	−	−	−	+	+	−	−	−	−	−	−	−	+	+	+	+	+	+	−	−
nasal	−	−	+	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	+	−
continuant	−	−	−	+	+	−	−	−	−	+	+	−	+	+	−	−	+	+	−	+
voiced	−	+	+	−	+	−	+	−	+	−	+	+	−	+	−	+	−	+	+	−

* Phonemes having the features −vocalic +consonantal.

The reasons for these particular choices are as follows. In general, we are interested in a consonant system in a single special environment, since it is clear that there are severe restrictions on the environments within which any given phoneme can occur (e.g. ŋ, the final phoneme in *sing*, never occurs at the beginning of a word). Therefore, it might be expected that there are differently organized systems of consonants for different environments, such differences being obscured if all environments are considered together. The initial decision to study English consonants at the beginning of words and immediately preceding long *a* was adopted since it was this environment that Miller & Nicely (1955) used in their very thorough study of acoustic confusions between consonants. The remaining choices follow from this decision, manipulating environment (end rather than beginning of word, as in case 2) or language (as in cases 4 and 5). Insufficient was known about Welsh vowels to attempt exactly the same analysis for Welsh; instead, consonants at the beginning of *all* Welsh words were studied. However, the rules governing all these systems are not very restrictive—a majority of the possible consonants can occur in every case—and in contrast a system where the rules impose severe restrictions on possible choices was studied (case 3—initial English consonants immediately preceding liquids).

In all these cases the technique of analysis was the same. The frequency of occurrence of each consonant in the environment in question was determined: 'frequency of occurrence' means number of *words* having this particular consonant in this particular environment, not number of *occurrences* of these words. This convention has been adopted since number of occurrences of words is very hard to determine (even Thorndike & Lorge is not complete in this respect), whereas number of words can be readily determined from a dictionary. In case 1, a check was performed on this procedure by showing that words beginning with a particular consonant did not interact with the frequency class to which they belonged, i.e. number of words and number of occurrences of words are equivalent measures of frequency.

When a word count such as Thorndike–Lorge is not available, small 'concise' dictionaries have been used, on the grounds that these will contain only the frequent words of a language, and rare words, which would be found in larger dictionaries and might not be typical of the language as it is commonly used, are thus omitted. Appendix 3 gives the results of this statistical procedure.

An empirical test of the cost models on these data can be achieved by observing response bias. Models 1, 2 and 3 all predict that it is less costly to be biased than unbiased for low discriminability, but more costly to be moderately biased than unbiased for high discriminability. Thus it is predicted that along poorly discriminable dimensions of the consonant system there would be more bias than would be expected by chance, but along highly discriminable dimensions there would be less than chance bias.

This hypothesis has been tested in the following way. First, from a given system, the number of consonants with a particular feature on a particular dimension has been counted (e.g. in case 1 there are 10 +diffuse consonants and five −diffuse consonants). Then the total number of occurrences of each of these types of consonant has been counted (in case 1 there are 386 +diffuse occurrences and 164 −diffuse occurrences). The following index is then calculated:

$$\text{Bias ratio} = \frac{\begin{array}{l}\text{Ratio of number of occurrences of consonants having a particular}\\ \text{feature to number of occurrences of consonants having the opposite}\\ \text{feature}\end{array}}{\begin{array}{l}\text{Ratio of number of consonants having a particular feature to}\\ \text{number of consonants having the opposite feature}\end{array}}$$

where both numerator and denominator are not less than 1; in this example

$$\text{bias ratio} = \frac{386}{164} \Big/ \frac{10}{5}.$$

Thus the bias ratio will be greater or less than unity according as the number of occurrences of consonants is more or less biased than the number of consonants, i.e. the bias ratio will indicate whether the bias of frequencies of occurrence actually observed is more or less than the bias that would be observed if the frequencies of occurrence were assigned to the possible consonants at random.

Table 4 shows the results of this procedure. There is a tendency for consonants with high random bias to have lower bias ratios than those with low random bias (Mann–Whitney *U*-test on high random bias (>2) and low random bias ($\leqslant 2$) groups, 2 per cent level, two-tailed). Thus doing separate Kruskal–Wallis one-way analyses of variance on the bias ratios of both groups, highly significant results are obtained ($H \simeq \chi^2 = 75$, 4 d.f., $P < 0{\cdot}001$ for low random bias group; $H \simeq \chi^2 = 52$, 3 d.f., $P < 0{\cdot}001$ for high random bias group).

Mean values of the bias ratio, corrected for the difference between the random bias groups, are as follows: grave, 1·70; diffuse, 1·325; strident, 2·35; nasal, 1·315; continuant, 2·31; voiced, 1·02. These values are computed by calculating separate mean bias ratios for the high random bias group and the low random bias group, then calculating for each dimension how much on average its bias ratios exceed the mean of the group each belongs to, and finally adding this value to the mean of the bias ratios of both random bias groups pooled together. This is not an entirely satisfactory procedure since the different random bias groups have quite different numbers of bias ratios from the various dimensions. However, it probably represents a fair first approximation.

This result can be related to the discriminability of the phonological features. Using Miller & Nicely's data, for frequency response 200–6500 c.p.s., and $S/N = -6$ db (there is nothing special about these physical parameters —this is simply the broadest sound spectrum Miller & Nicely used, with a moderately difficult white-noise level, enough to produce confusions, but not excessively so), η, the choice theory measure of discriminability, takes the following values: grave, 2·33; diffuse, 2·86; strident, 2·56; nasal, 31·3; continuant, 5·00; voiced, 33·3; i.e. descending order of discriminability is voiced, nasal, continuant, diffuse, strident, grave.

There is a marked correlation between order of bias ratio and order of discriminability. The rank orders of the bias ratios from each of the six samples, corrected in the manner described above, agree well with each other (Kendall $W = 0{\cdot}55$, $P < 0{\cdot}001$) and the rank order of the bias ratios from each sample correlates positively with the order of discriminability derived from Miller & Nicely (1955), mean Kendall correlation for the six samples $= 0{\cdot}44$, significantly different from zero (*t*-test, $P < 0{\cdot}001$, two-tailed).

The present hypothesis would predict an almost perfect correlation between these two measures: failure to achieve this perfect correlation could be explained by the unexpectedly low bias ratios of the two place-of-articulation features, grave and diffuse. Now the cost models are most closely concerned with the receiver's discriminability problems, and the transmitter's articulation problems have not been considered; and these clearly need incorporating in any complete

TABLE 4. BIAS RATIOS

	Ratio of number of consonants: ‘ random bias ’	Ratio of number of occurrences of consonants: ‘ actual bias ’	Bias ratio
CASE 1			
grave	7:8	432:118	3·20
diffuse	10:5	386:164	1·18
strident	6:9	110:440	2·67
nasal	2:13	104:446	0·66
continuant	5:10	83:467	2·81
voiced	7:8	235:315	1·17
CASE 2			
grave	6:10	79:112	0·85
diffuse	11:5	142:49	1·32
strident	8:8	60:131	2·18
nasal	2:14	27:164	0·87
continuant	7:9	53:148	2·17
voiced	8:8	85:106	1·25
CASE 3			
grave	10:6	1622:474	2·15
diffuse	10:6	1501:595	1·51
strident	5:11	372:1724	2·11
continuant	6:10	410:1686	2·47
voiced	5:11	638:1458	1·03
CASE 4			
grave	7:7	2288:798	2·87
diffuse	10:4	2195:891	0·99
strident	6:8	750:2336	2·34
nasal	2:12	613:2473	0·67
continuant	6:8	750:2336	2·34
voiced	8:6	1340:1746	0·98
CASE 5			
grave	7:5	421:259	1·16
diffuse	9:3	511:169	1·01
strident	4:8	176:504	1·43
nasal	2:10	167:513	0·61
continuant	4:8	176:504	1·43
voiced	7:5	419:261	1·15
CASE 6			
grave	8:6	4168:2149	1·45
diffuse	10:4	3812:2505	0·61
strident	4:10	666:5651	3·39
nasal	2:12	603:5714	1·57
continuant	5:9	769:5548	4·01
voiced	7:7	2818:3499	1·24

cost treatment. (There is an old study of Sacia & Beck, 1926, reported by Miller, 1951, where energy involved in articulating a consonant and its frequency of occurrence are negatively correlated.) It is possible that these latter considerations will have some influence on the distribution of features, and the grave and diffuse dimensions, being most associated with the largest physical changes in the articulation process, might be most affected.

However, a case has been made out for at least some contribution of discriminability and cost to the distribution of consonants in a number of different environments and in a number of different languages. It is suggested that such factors have shaped the statistical structure of these languages over the course of time; i.e. the interaction between discriminability and cost has led speakers of the language to adjust their response biases for words of different phonetic structures, and this adjustment is reflected in the frequency of usage of differently structured words.

5. SUMMARY

This paper is preliminary: preliminary both in the sense that it precedes papers that will give more detailed applications of the cost models presented here; and in the sense that the broad definition of cost that was given in the introduction has not been followed by a broad range of models. In restricting the discussion to the behaviour of the 'ideal observer' of signal detectability theory in very limited conditions, the intention has been to lay a foundation on which ultimately models of cost can be built having sufficient elaboration to satisfy, for example, the psycholinguist or the control theorist. However, some successes have been achieved, even within the terms of reference of this paper: (1) the device of taking some of the procedures of signal detectability theory literally (namely, that an observer must calculate a likelihood ratio and determine whether this value exceeds a previously determined criterion level) has produced models that show important and slightly unexpected interactions between discriminability and response bias; (2) the similarity between models 2 and 3 means that signal detectability or choice theory concepts can be used interchangeably in this sort of cost model; (3) some successes can be claimed for model 2 in its analysis of such diverse topics as reaction time, vigilance and linguistic studies. This, of course, is the value of the concept of cost: by definition, it must be widely applicable.

ACKNOWLEDGEMENTS

This work was carried out partly while the author held an M.R.C. scholarship at the M.R.C. Psycholinguistics Unit, Oxford, and partly while he was Clifford Norton Junior Research Fellow at the Queen's College, Oxford. The author wishes to thank Professor R. J. Audley for help at many stages in the preparation of this paper. The treatment in Section 3 was improved by the suggestions of a referee.

REFERENCES

ABRAMSON, N. (1963). *Information Theory and Coding.* New York: McGraw-Hill.

AUDLEY, R. J. (1960). A stochastic model for individual choice behaviour. *Psychol. Rev.* **67,** 1–15.

BERTELSON, P. (1961). Sequential redundancy and speed in a serial two-choice responding task. *Q. J. exp. Psychol.* **13,** 90–92.

BROADBENT, D. E. & GREGORY, M. (1963). Vigilance considered as a statistical decision. *Br. J. Psychol.* **54,** 309–323.

CARTERETTE, E. C., FRIEDMAN, M. P. & COSMIDES, R. (1965). Reaction time distributions in the detection of weak signals in noise. *J. acoust. Soc. Am.* **38,** 531–542.

CHOMSKY, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, Mass.: M.I.T. Press.

GARNER, W. R. (1962). *Uncertainty and Structure as Psychological Concepts.* New York: Wiley.

GREEN, D. M. (1960). Psychoacoustics and detection theory. *J. acoust. Soc. Am.* **32,** 1189–1203.

HALLE, M. (1958). Questions of Linguistics. *Nuovo Cimento* (Suppl.) **13,** 494–517. Revised version (1964) entitled ' On the Bases of Phonology ', in J. A. Fodor & J. J. Katz (eds.), *The Structure of Language.* Englewood Cliffs, N.J.: Prentice-Hall.

HALLE, M. (1965). *Introduction to Linguistics.* Lecture course given at M.I.T.

HUFFMAN, D. A. (1952). A method for the construction of minimum redundancy codes. *Proc. I.R.E.* **40,** 1098–1101.

JAKOBSON, R. (1953). *Essais de linguistique générale.* Paris.

JAMES, W. (1890). *The Principles of Psychology*, vol. 1, ch. 4. New York: Holt.

LUCE, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. 1, pp. 103–189. New York: Wiley.

MILLER, G. A. (1951). *Language and Communication.* New York: McGraw-Hill.

MILLER, G. A. & NICELY, P. E. (1955). An analysis of perceptual confusions among some English consonants. *J. acoust. Soc. Am.* **27,** 338–352.

SACIA, C. F. & BECK, C. J. (1926). The power of fundamental speech sounds. *Bell Syst. Tech. J.* **5,** 393–403.

SHANNON, C. E. (1948). The mathematical theory of communication. *Bell Syst. Tech. J.* **27,** 379–423.

STONE, M. (1960). Models for choice reaction time. *Psychometrika* **25,** 251–260.

SWETS, J. A. & GREEN, D. M. (1961). Sequential observations by human observers of signals in noise. In E. C. Cherry (ed.), *Proc. 4th London Symp. Information Theory,* pp. 177–195. London: Butterworths.

THORNDIKE, E. L. & LORGE, I. (1944). *The Teacher's Word Book of* 30,000 *Words.* New York: Teachers' College, Columbia University.

THURSTONE, L. L. (1930). The learning function. *J. gen. Psychol.* **3,** 469–493.

ZIPF, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, Mass.: Addison-Wesley.

APPENDIX 1

Central Dip in the Cost Function for Large d'

y, the sum of the probability density functions of two eqiprobable normally distributed stimuli is proportional to $\exp(-\frac{1}{2}x^2) + \exp[-\frac{1}{2}(d'-x)^2]$. So that

$$dy/dx \propto -x \exp(-\tfrac{1}{2}x^2) - (x-d') \exp[-\tfrac{1}{2}(d'-x)^2]$$

and

$$d^2y/dx^2 \propto -\exp(-\tfrac{1}{2}x^2) + x^2 \exp(-\tfrac{1}{2}x^2) - \exp[-\tfrac{1}{2}(d'-x)^2] + (x-d')^2 \exp[-\tfrac{1}{2}(d'-x)^2].$$

At the central criterion level, $x = \frac{1}{2}d'$,

$$dy/dx = 0 \quad \text{and} \quad d^2y/dx^2 \propto \exp(-\tfrac{1}{8}d'^2)[\tfrac{1}{2}d'^2 - 2],$$

which is $\gtreqless 0$ as $d' \gtreqless 2$. Therefore $x = \frac{1}{2}d'$ is a maximum if $d' < 2$ and a minimum if $d' > 2$, i.e. the central dip in the cost function appears only for $d' > 2$. Q.E.D.

APPENDIX 2

Derivation of Choice Theory Results in the Case of Two Equiprobable Stimuli

Suppose $\eta(s, s) = \eta(s', s')$ and write

$$\frac{\eta(s, s)}{\eta(s, s')} = \eta \quad \text{and} \quad \frac{b(r)}{b(r')} = b.$$

Then

$$p(r|s) = \frac{\eta b}{\eta b + 1},$$

$$p(r'|s) = \frac{1}{\eta b + 1},$$

$$p(r|s') = \frac{b}{\eta + b},$$

$$p(r'|s') = \frac{\eta}{\eta + b},$$

$$p(s) = p(s') = \tfrac{1}{2}.$$

Mean cost $= \bar{E}$

$$= -\sum_s p(s) \sum_r p(r) \log p(r|s)$$

$$= \tfrac{1}{2}\left\{\frac{\eta b}{\eta b + 1}[\log(\eta b + 1) - \log \eta b] + \frac{1}{\eta b + 1}[\log(\eta b + 1)]\right\} + \tfrac{1}{2}\left\{\frac{b}{\eta + b}[\log(\eta + b) - \log b] + \frac{\eta}{\eta + b}[\log(\eta + b) - \log \eta]\right\},$$

$$2\bar{E}=\log(\eta b+1)+\log(\eta+b)-\eta\log\eta\left[\frac{1+2\eta b+b^2}{(\eta b+1)(\eta+b)}\right]$$
$$-b\log b\left[\frac{1+2\eta b+\eta^2}{(\eta b+1)(\eta+b)}\right]$$

$$2(\eta b+1)(\eta+b)\bar{E}=(\eta b+1)(\eta+b)[\log(\eta b+1)+\log(\eta+b)]$$
$$-\eta\log\eta[1+2\eta b+b^2]-b\log b[1+2\eta b+\eta^2].$$

Differentiating with respect to b,

$$2[1+2\eta b+\eta^2]\bar{E}+2(\eta b+1)(\eta+b)\frac{d\bar{E}}{db}$$
$$=[1+2\eta b+\eta^2][\log(\eta b+1)+\log(\eta+b)]$$
$$+(\eta+b)\eta+(\eta b+1)-2\eta\log\eta[\eta+b]$$
$$-b\log b[2\eta]-[1+\log b][1+2\eta b+\eta^2].$$

Differentiating with respect to b again, putting $b=1$ and using the fact that $d\bar{E}/db=0$,

$$4\eta\bar{E}+2(\eta+1)^2\frac{d^2E}{db^2}=[1+\eta^2]\left[\frac{\eta}{\eta+1}+\frac{1}{\eta+1}\right]$$
$$+2\eta[2\log(\eta+1)+1]-2\eta\log\eta-4\eta-[1+\eta]^2.$$

When $b=1$,

$$2\bar{E}=2\log(\eta+1)-\frac{2\eta}{\eta+1}\log\eta.$$

Thus

$$(\eta+1)^2\frac{d^2\bar{E}}{db^2}=\eta\left[2\log(\eta+1)-\log\eta-1-2\log(\eta+1)+\frac{2\eta}{\eta+1}\log\eta\right]$$
$$=\eta\left[\frac{\eta-1}{\eta+1}\log\eta-1\right].$$

Therefore $\bar{E}$ is a maximum or minimum at $b=1$ according as $d^2\bar{E}/db^2$ is negative or positive, which is according as

$$\frac{\eta-1}{\eta+1}\log\eta\lesseqgtr 1.$$

The solution to

$$\frac{\eta-1}{\eta+1}\log\eta=1$$

is $\eta=4{\cdot}681$.

$\bar{E}$ is a $\left\{\begin{matrix}\text{maximum}\\ \text{minimum}\end{matrix}\right\}$ at $b=1$ according as $\eta\lesseqgtr 4{\cdot}681$. Q.E.D.

APPENDIX 3

Frequency Distributions of Various Consonant Systems

Case 1: English consonants C of the form #Cā.
Case 2: English consonants C of the form āC#.
Case 3: English consonants C in the environment #Cl or #Cr.
Case 4: French consonants C in the environment #Ca.
Case 5: German consonants C in the environment #Ca.
Case 6: Welsh consonants C in the environment #C.
(# = word boundary.)

Case 1		Case 2		Case 3		Case 4		Case 5		Case 6	
Consonant	Frequency (no. of words)	Consonant	Frequency (no. of words)	Phoneme pair	No. of words	Consonant	No. of words	Consonant	No. of words	Consonant	No. of words
p	104	d	37	pr	393	k	532	k	107	k	1473
m	98	k	31	tr	227	m	520	m	85	d	946
k	92	t	23	gr	189	p	479	n	82	g	922
b	59	m	20	br	188	b	297	d	61	p	676
f	37	s	18	kr	181	s	249	v	60	t	606
g	30	f	15	kl	140	t	204	b	56	m	455
d	25	ǰ	7	fr	136	g	160	p	47	s	438
t	25	n	7	fl	127	f	155	z	46	b	318
č	23	č	6	pl	107	v	145	f	37	f	199
s	19	v	6	bl	100	š	127	t	37	n	148
š	14	θ	6	dr	100	n	93	š	33	x	103
v	12	p	5	sl	85	ž	72	g	29	v	22
n	6	ž	3	gl	61	d	51			ǰ	7
ǰ	5	z	3	θr	38	z	2			θ	4
θ	1	š	2	šr	23						
		b	2	šl	1						

Vol. 21
Part 1
61–84

The British Journal of Mathematical and
Statistical Psychology

May
1968

THE USE OF A NULL-POINT METHOD TO STUDY THE ACQUISITION OF SIMPLE AND COMPLEX TRANSFORMATION SKILLS

By GORDON PASK and B. N. LEWIS[1]

System Research Ltd., Richmond, Surrey

A null-point experimental method for measuring human performance is described. This method can be used in studies of learning and in other situations where the effect of learning is appreciable, though irrelevant to the experimenter's objective. It provides a form of constancy of experimental conditions which is appropriate if the subject is conceived as an imperfect goal-directed system. The method is exhibited in connexion with two experiments and a subsidiary study. One experiment involves the acquisition of a skill in which the human subject learns to apply a single transformation rule. For the other main experiment, the subject learns to deal alternately with two rules (so chosen that interference as well as positive transfer of training takes place). Further, in this case, the subject learns a concept which (in an operational sense) comprehends the pair of rules. The subsidiary experiment involves a modification of the original technique which may be useful in more detailed investigations of concept learning.

1. INTRODUCTION

This paper describes a null-point experimental method and its application in a series of human learning experiments. The method is shown diagrammatically in Fig. 1. The subject is learning a skill that entails the solution of a sequence of problems (in the cases considered, these are coding problems, but the method is quite general). The control mechanism, M, instructs the subject in the skill by adjusting the problem difficulty as a function of his behaviour, so that the subject's performance is held at a fixed level, the system null-point.

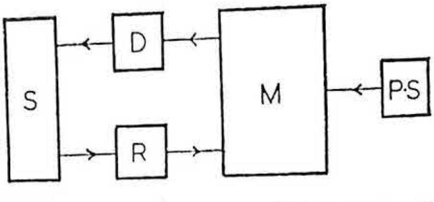


FIGURE 1. Null-point or steady-state system. M changes the difficulty to compensate for learning in the subject. The coupled system 'subject, M' does not learn. S, subject; M, control mechanism; D, display; R, response; $P \cdot S$, sequence of problems typical of the skill.

---

[1] Now with the University of London Institute of Education.

M's of this type have been extensively used as teaching devices (Pask, 1958 *a*, *b*, 1960, 1961 *a*, *b*, 1963 *a*, *b*, 1964; Lewis & Pask, 1965) and in this capacity *M* may be designed to mediate a more refined adjustment procedure; however, for a certain class of skills, the maintenance of a suitable null-point is also quite a good teaching strategy. For the present purpose, it is not necessary to examine the teaching function of *M* explicitly.

The operation of *M* is, of course, constrained in various ways. In the first place, *M* must not overload the subject by presenting problems so difficult that they are unintelligible. On the other hand, *M* has to provide sufficient relevant novel experience to sustain the subject's attention. It has been argued in other papers (Pask, 1961 *a*, *b*, 1966) that *M* has to satisfy a number of more subtle requirements as well as these.

The experimental advantage that is gained by using an *M* stabilized system is a standard condition for measurement. The progress of learning can, for example, be expressed in terms of the difficulty of problem that the subject is able to solve at a constant level of proficiency. The ' constant proficiency ' measurement is often useful, as it stands, in ergonomics and design studies (Hudson, 1963; Kelly, 1966); it is a potent type of measurement when ' difficulty ' (or its converse variate ' simplification ') can be interpreted within the framework of a problem-solving or learning model (Pask, 1964, 1966), whereas the detailed response mechanism cannot be represented in this fashion.

Most importantly, if we entertain the cybernetic concept that man is a goal-directed system, then a technique of this sort provides the appropriate condition for measuring his performance. For, if man is a goal-directed system at all, he is usually an imperfect one. His imperfections may be due to inattention or fatigue or an inadequate repertoire of mental processes. In any case, they are often remedied by learning. The null-point method offers the possibility of studying man in conditions that ensure constant *efficiency* in goal achievement. If man is a goal-directed system, then it is *this* constancy, rather than a constancy of stimulus conditions, which ought to be preserved from experiment to experiment.

Several other standard conditions may be maintained by *M* stabilization methods (Verhave, 1963; Pringhorn, 1961; Pask, 1963 *a*, *b*). In each case, the experimenter replaces examination of a subject who (by definition) has non-stationary characteristics that thwart individual measurement, by the examination of a coupled ' subject, *M* ' system (as in Fig. 1). The coupled ' subject, *M* ' system, regarded as a whole, does *not* learn; it is a stable dynamic system and its individual behaviour can be observed and measured without difficulty.

In principle, the experimenter could stand in the place of *M*. In practice, some sort of mechanization is essential. The various *M*'s used in the present experiments were simulated on a special-purpose computer that controlled the subject's display and received an input from his response board.

In Experiment 1, *M* controlled the acquisition of a simple code-learning skill. In Experiment 2, *M* controlled a more complex situation where the subject was required to deal with the alternation of two different codes: on any trial one code is named and used, but the subject is unable to predict which code this will be beforehand. The latter situation makes the subject divide his attention between the two codes, or the cognitive processes responsible for dealing with them, and the rapid and unpredictable alternation gives rise to a great deal of interference between the sub-skills of applying the separate codes even though, owing to code similarity, there is some positive transfer of training. Hence, it is possible to study the development of a conceptual structure that enables the subject to treat the complex skill as a whole and thus to overcome the effects of interference.

Finally, in both Experiments 1 and 2, it was possible to observe an interesting mode of man–machine interaction (called 'participant interaction' because it is symptomatic of an attempt, on the subject's part, to control the strategy of M). Briefly, the stimulus response communication between the subject and M can be viewed as discourse in a zero-level object language (designated L^0). The vocabulary of L^0 consists in signs for stimuli and responses and its syntax is a set of rules for constructing legal responses. However, L^0 contains no legal expressions that M can interpret as instructions to change its strategy or the form of stimuli it presents. All the same, a subject can learn about the control principles built into M and he can *misuse* L^0 in order to modify the experimental situation. He can, for example, make responses that M will interpret as mistakes (even though the subject is able to solve the problem posed by the stimulus and is not really mistaken when he makes such a response). As a result of this behaviour, M reacts as though the subject had actually been mistaken (by reducing the difficulty of the task, or, in a code alternation skill, in a more elaborate fashion).

Various instances of 'participant interaction' have been observed in the conduct of the experiments. Some attempts have been made to investigate the phenomenon and to provide higher levels of communication (object languages, L^1) that bring this mode of discourse into the open.

2. EXPERIMENT 1

The object of Experiment 1 was to realize a 'subject, M' system in the specific milieu of a code application or transformation skill (without code alternation) and to investigate the stability of different 'subject, M' systems.

The transformation skill has the following characteristics. A subject is presented with an unlearnable sequence of stimuli spaced equally in time taking the form of configurations of up to four illuminated signal lamps (selected from a group of eight lamps). Before the experiment begins, he is told that a transformation rule, Ω, relates the group of signal lamps to a row of eight response buttons; he knows this rule and is able to recite it. Given a stimulus (designated x) he is required to solve the problem of producing a response, $y=\Omega(x)$, within a deliberately restricted interval, δt (his response is complex, for he must press several buttons to denote the transformation of a stimulus configuration).

The size of the stimulus configuration, the value of δt and the inter-trial interval are chosen so that (1) a fully proficient subject can perform the skill accurately, but (2) a novice finds the problems posed by x in the context of Ω unsolvable (e.g. his uncertainty regarding the value of y when x is given is such that he is grossly overloaded).

In these conditions a novice can neither perform the skill nor learn to perform it unless he is provided with external assistance. If a marginal amount of external assistance is provided by M, then the subject is forced to exhibit various problem-solving behaviours which would, in more relaxed circumstances, be unobservable.

In order to start a learning process, M must reduce the difficulty of the situation. It does so by 'simplifying' or 'partially solving' the problems posed by the stimuli. A combination of two simplifying procedures is used for this purpose.

One procedure consists in reducing the number of lamps in a stimulus configuration. The other consists in providing cue information that partially specifies the correct response. Both of these procedures, as well as their combination (to form a conjoint simplifying operation), have been validated (1) with respect to a computer-programmed learning model in which they serve to solve partially the problems posed by stimuli, x, and (2) by checking, empirically, that the percentage of correct responses elicited by a sequence of more simplified stimuli is never less than the percentage of correct responses elicited by a sequence of *less* simple stimuli (and that, if we exclude the case of a fully proficient subject, it is greater). The snare to be avoided is that some intuitively acceptable simplifying procedures suffer from an 'inversion' noticed by van der Veldt (cited in Woodworth, 1950). If a subject is taught to respond to a complex stimulus by way of simple exemplars, the subsequent presentation of a less complex stimulus may give rise to more, rather than less difficulty; the subject has acquired a perceptual motor organization applicable to the complex situation; the organization apposite to less complex situations has decayed.

During the learning process, M computes an average correct response rate (called 'proficiency' and designated ρ). It continually adjusts the degree of simplification (designated μ) so that the subject's proficiency is held constant (ρ tends to a fixed value ρ_0). In the terminal condition, when the subject has learned the skill, the degree of simplification is reduced to zero. After this point, further learning leads to an uncontrolled increase in the value of ρ.

The crux of these comments is summarized in the predicted learning curve of Fig. 2. For graphical convenience, the 'degree of simplification', μ, has been transformed to an index of 'objective difficulty' $\eta = \mu_{\max} - \mu + 1$. Phase I of the curve is a transient condition, in which M is seeking a value of η that fixes ρ at ρ_0 for the naive subject. Phase II represents the controlled part of the learning process where η is increased to maintain ρ constant. Phase III shows the uncontrolled increase in ρ for η held at its maximum value (degree of simplification is zero). Finally, Phase IV is an unpredictable part of the curve in which, according to his motivation, the subject may either stick at the job or succumb to boredom or fatigue.

The chosen value of ρ_0 influences the subjective situation of the learner and may be expected to modify the objective efficiency of his learning. If ρ_0 is low in value, the subjective situation is ' stringent ' in the sense that a small increase in proficiency leads to an increase in the difficulty of the task. Conversely, if ρ_0 is high, the situation is relaxed. Hence, for low values of ρ_0 the subject has to deal with more difficult problems and his uncertainty (regarding its solution) is greater, on average, than it would be for a high value of ρ_0.

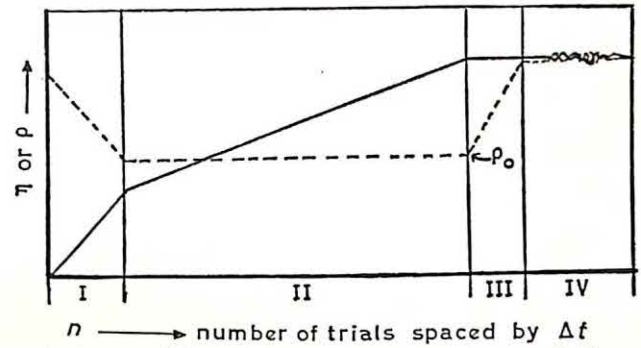


FIGURE 2. Predicted form of learning curve in ' subject, *M* ' system. - - - - -, ρ; ———, η. Phase I: Initial transient; η increases; ρ approaches $\rho=\rho_0$. Phase II: Steady state; $\rho \rightarrow \rho_0$; η increases. Phase III: Uncontrolled increase in ρ with $\eta=\eta_{\max}$. Phase IV: Unpredictable part.

Various predictions can be made about the effect of ρ_0 upon objective indices of learning. In fact, we advanced the hypothesis that learning of this type relies upon an information processing mechanism (a functional entity, perhaps to be identified with ' intermediate ' or ' constructive ' memory) and that this mechanism works best when (1) it is maximally loaded with relevant data, providing that (2) it is not overloaded. This hypothesis leads to the prediction that the learning rate will *increase* with *decreasing* ρ_0 provided that ρ_0 is not forced below a minimum value.

Method

The different ' subject, *M* ' systems in Experiment 1 were characterized by simple and readily instrumented control rules. The control rules all had the effect of increasing η or decreasing μ as the subject's measured proficiency increased. They differed in so far as different values were assigned to a parameter, ξ (the increment in proficiency required to bring about a unit decrease in simplification).

The level of proficiency, ρ_0, at which the subject is held (throughout most of the training process) depends upon the value of ξ; and different control rules may, equivalently, be characterized by different mean values of ρ_0. The dependent variable in Experiment 1 was an index of the learning rate for groups of subjects assigned to systems having different values of ξ, different resulting values of ρ_0 and different 'stringencies'.

Apparatus

Display and Response Arrangements The display and response board used in Experiment 1 comprised a group, A, of eight signal lamps, lettered *a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, which were used to deliver stimuli; a row, B, of eight lamps providing knowledge of results data and a modicum of cueing information; a row, C, of eight lamps that form part of an external 'memory' facility and a row, D, of response buttons, numbered, 1, 2, 3, 4, 5, 6, 7, 8. The members of rows B, C and D were in one-to-one correspondence and separated by a vertical distance of one inch. The lamps in group A were arranged in two rows of four and were situated 11 inches above row B.

Stimulus Presentation In the unsimplified case (obtaining when the subject is proficient) the stimuli, x, consisted of configurations of four simultaneously illuminated lamps, selected out of group A. The stimulus lamps remained on for an interval of δt (which, in Experiment 1, was set at 4 sec.).

The apparatus was designed to present eight possible unsimplified stimuli which are the following lamp configurations: (I)$=abch$, (II)$=abdc$, (III)$=bdce$, (IV)$=cedf$, (V)$=egdf$, (VI)$=fheg$, (VII)$=gafh$ and (VIII)$=gahb$. Simplified stimuli are configurations of m lamps (m less than or equal to four) which are subsets of these configurations and which may be accompanied with cueing information delivered by simultaneously illuminating certain lamps in row B. Hence, any simplified stimulus is derived from an unsimplified stimulus selected for presentation at a given trial.

Unsimplified stimuli were selected by a quasi-'random' apparatus at intervals of Δt (in Experiment 1, $\Delta t = 5{\cdot}5$ sec.). Each of the eight possible stimuli appeared with equal frequency in the stimulus sequence over a block of 24 trials; apart from this constraint, the sequence is arbitrary and is intended to be unlearnable.

Response System A mapping, Ω (experimentally interpreted as a transformation rule), relates the elements denoted by the lamps in row B or (since the rows are in one-to-one correspondence) to the lamps in row C or the response buttons in row D. In Experiment 1, Ω was the stimulus response permutation *a*, 2; b, 1; *c*, 6; *d*, 5; *e*, 4; *f*, 3; *g*, 8; *h*, 7.

The subject responded by selecting a configuration, y, of response buttons and he was *required* to select the particular configuration $y = \Omega(x)$. Hence, if the stimulus at the nth trial (designated $x(n)$) has m components, then a *correct* response, $y(n) = \Omega[x(n)]$ also has m components. Further, in order to count as a correct response, $y(n)$ must be completed within the interval δt during which $x(n)$ is exposed; in other words δt serves as an interval allowed for response (and the response buttons are automatically rendered ineffective after δt has elapsed).

However, the response buttons need not be pressed simultaneously and the subject is permitted to perform the construction of y in any order he pleases. To emphasize these conditions and to avoid an unwanted load on his immediate memory, each response button was associated with a ' lock ' relay; this relay is energized if the corresponding button is pressed and it remains energized until 0·5 sec. after the end of δt. Whilst the relay is energized, the corresponding external memory lamp in row C is illuminated to provide a visual record of y. The interval δt is terminated by the disappearance of the stimulus lamps and the coincident illumination of those ' knowledge of results ' lamps in row B that indicate the correct response $\Omega(x)$. This presentation persists for 0·5 sec. after the end of δt (so that it coincides with the image of y, thus allowing the subject to compare $\Omega(x)$ and his actual response, y). Coincidentally, the subject receives an auditory and visual signal (buzzer and lamp) to indicate whether his response is completely correct (all response selections correct) or mistaken. At an instant, $\delta t+0{\cdot}5$ sec. after the appearance of the stimulus, the ' lock ' relays are de-energized and the lamps in rows B and C extinguished. There is a rest interval of 1 sec. and the next stimulus is presented $\Delta t=5{\cdot}5$ sec. $=\delta t+0{\cdot}5+1$ sec. after the last.

Simplification Procedures M, schematized in Fig. 3, includes equipment for mediating two simplification procedures. One of these involves a variation in m (the number of lamps in a stimulus configuration) and the other involves the provision of cue information if, and only if, a binary ' cue variable ', z, is equal to 1.

The first procedure relies upon the mechanical embodiment of the following hierarchical array of subsets of the unsimplified (or $m=4$) stimulus configurations (which form the upper row in the hierarchy).

m	I	II	III	IV	V	VI	VII	VIII
4	*abch*	*abdc*	*bdce*	*cedf*	*egdf*	*fheg*	*gafh*	*gahb*
3	*abc*	*bdc*	*dce*	*def*	*efg*	*fgh*	*hga*	*ahb*
2	*ac*	*bd*	*ce*	*df*	*eg*	*fh*	*ga*	*hb*
1	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*

At any trial, the equipment receives an instruction (1) to select a basic stimulus configuration out of the eight possibilities (labelled here with Roman numerals) and (2) to select this stimulus with a given value of m. Part (1) of the instruction determines a column in the array and part (2) determines a row. The element at the row-column intersection is displayed in lamp group A.

The other simplifying procedure consists in the provision of cue information, through the row B lamps, *within* the interval, δt, in which a response can be submitted and accepted. Cue information partially specifies a correct response, $y=\Omega(x)$. As above, it is provided if a variable $z=1$ and is not provided if $z=0$. Specifically, if $z=1$ and if x is an m-lamp configuration, then *after* the subject has selected $m-1$ correct members of y, he is automatically provided with an image (through the row B lamps) of the remaining selection that he must make in order to achieve a correct response.

These procedures are combined to provide eight degrees of simplification indexed by the variable, μ (or its converse variate, $\eta=\mu_{\max}-\mu+1$). The combination of procedures is determined by the listing shown below:

Value of μ	0	1	2	3	4	5	6	7
Value of η	8	7	6	5	4	3	2	1
Value of m	4	4	3	3	2	2	1	1
Value of z	0	1	0	1	0	1	0	1

Hence, a given value of η or μ (provided at each trial by the output of a decision rule in M) generates a unique value of m and a unique value of z. Further, at each trial, the 'random' stimulus generator selects a basic (Roman numeral) stimulus configuration. This selection, modified according to the prevailing value of η or μ, determines the stimulus presented to the subject.

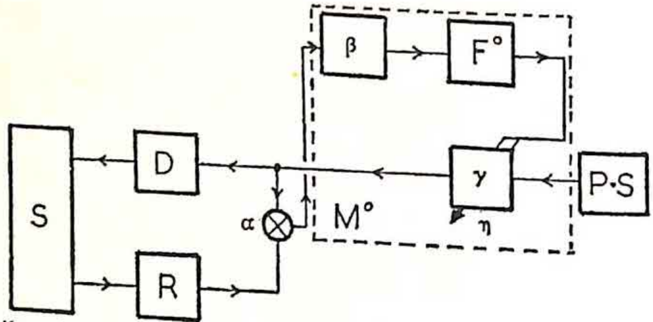


FIGURE 3. Details of control system. S, subject; M, control mechanism; D, display; $P \cdot S$, sequence of problems; R, response; F^0, control rule; α, comparator producing value of $\sigma(n)$; β, averager producing value of $\rho(n)$; γ, simplification procedure.

Proficiency Measures Given $x(n)$, the rule Ω determines the value of a correct response. The least elaborate proficiency measure, ρ, and mean proficiency measure, $\bar{\rho}$, are obtained by averaging the value of a correct response function:

$$\begin{aligned}\sigma(n)&=1 \quad \text{if} \quad y(n)=\Omega[x(n)]\\ &=0 \quad \text{if not}\end{aligned}$$

over a fixed number of trials, say τ trials. Thus the incremental proficiency at trial $n=n_0$ (used as a control variable) is

$$\rho(n_0)=1/\tau \sum_{n=n_0-\tau+1}^{n=n_0} \sigma(n),$$

where (for the present experiment) τ is taken as 8. The proficiency measure $\bar{\rho}$ is defined as the 'long-term' average of ρ; in the data that are presented, the 'long-term' average has been taken over 24 trials.

Control Rule As shown in Fig. 2, M includes a decision rule or control rule which accepts the current value of $\rho(n)$ and determines the value of $\mu(n)$ or its converse variate $\eta(n)$. For Experiment 1, M embodies the control rule F^0 shown below:

$$\Delta\eta(n) = +1 \quad \text{if} \quad \rho(n) > \xi$$
$$\text{unless} \quad \eta(n) = \eta_{\max} = 8 \quad \text{when} \quad \Delta\eta(n) = 0;$$
$$\Delta\eta(n) = -1 \quad \text{if} \quad \xi^* > \rho(n)$$
$$\text{unless} \quad \eta(n) = \eta_{\min} = 1 \quad \text{when} \quad \Delta\eta(n) = 0;$$
$$\Delta\eta(n) = 0 \quad \text{if} \quad \xi > \rho(n) > \xi^*.$$

Initial condition; that $\eta(0) = 1$; $\xi^* = \xi - 0{\cdot}15$.

The value of ξ (the parameter varied in the experiment) determines the stringency of the situation, since it specifies the increase in ρ required to produce a unit increment in η.

Subjects The 30 subjects used in this experiment were young men and women from various backgrounds (about half of them being students and half being either Mensa members or local people). They were all familiar with mechanized situations, since they had acted as subjects in this laboratory, for other experiments.

Procedure As a preliminary treatment, each subject was introduced to the apparatus and the skill and he was rehearsed on an irrelevant transformation rule. After this, but before the main experiment, each subject was required to memorize Ω and to recapitulate it: (1) by drawing the mapping without any mistakes; and (2) by pairing lettered stimulus elements with numbered response elements.

As a result of the preliminary treatment, the subjects knew what to expect from M, but the character of the control mechanism was briefly discussed and any questions about it were answered.

Finally, the subject returned to the apparatus and was told that he must learn to apply Ω successfully to unsimplified stimuli, e.g. in a terminal condition of $\eta = \eta_{\max} = 8$.

Groups of 10 subjects each were assigned to three parametrically different control rules: $\xi = 0{\cdot}55$, $0{\cdot}65$ and $0{\cdot}75$. (These values of ξ had been found workable in pilot experiments and they yielded ρ_0 values in the regions 0·5–0·6, 0·6–0·7 and 0·7–0·8, respectively.) The value of ρ and η was recorded for each subject at each trial in the experiment to which he was assigned.

In order to compare learning in each condition, it is necessary to define a terminal criterion of proficiency. This criterion was taken as:

$$\eta(n) = \eta_{\max} = 8 \quad \text{and} \quad \rho(n) > 0{\cdot}75$$

for $2\tau = 16$ consecutive trials. The number of trials required to reach this criterion is designated T and T values were recorded for each subject. Since Δt is constant, T is also an index of the time required to reach a criterial level of performance. To eliminate the possibility of a subsequent regression in performance, the experiments were run until the subject achieved a criterion of $\rho(n) > 0{\cdot}8$, again for 2τ consecutive trials (in fact, no regression occurred). At this point the subject was told that his performance was satisfactory and he was dismissed.

The value of T slightly overestimates the number of trials (or the time) taken to pass through phase I and phase II of the learning curve shown in Fig. 2. Since there is a rather rapid increase in ρ (from its steady-state value of ρ_0) once that η consistently equals $\eta_{\max}$, the estimate is not *widely* different in form for the different values of ρ_0.

Results

The behaviour of all the subjects (in each of the three conditions) conformed to the paradigm of Fig. 2. A typical learning curve (man–machine system trajectory) is shown in Fig. 4. It is for a subject who was assigned to the $\xi=0{\cdot}55$ condition and who attained a ρ_0 value of between 0·5 and 0·6. In the region that corresponds to phase II of Fig. 2 these trajectories are *stable* in so far as perturbations in ρ are compensated by variations in η.

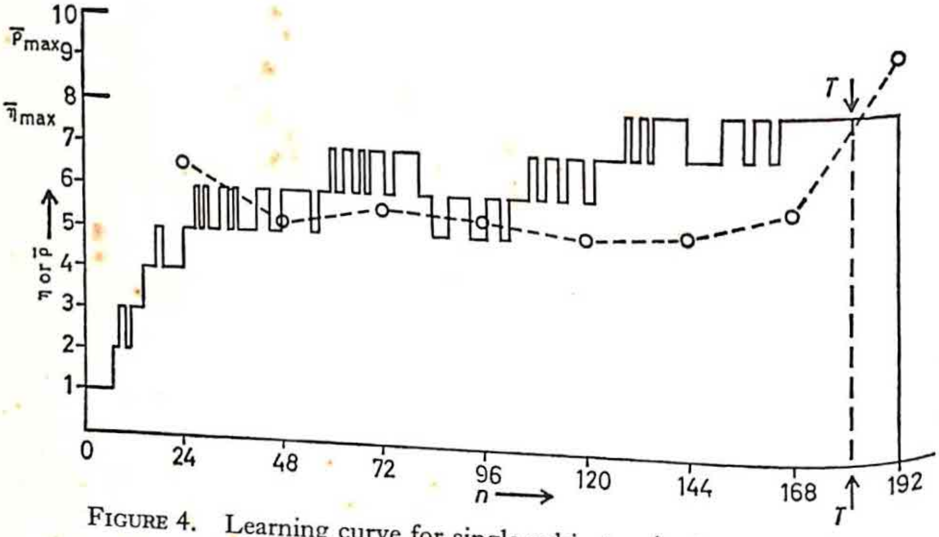


FIGURE 4. Learning curve for single subject. $\xi=0{\cdot}55$; $T=181$.

System stability is, of course, contingent upon certain limits which have been respected in these experiments. During the pilot experiments, used to determine workable values of ξ, it was found that the limits depended, quite crucially, upon the operating levels of ξ and ρ_0.

The main data consist in T values for each condition in the experiment; namely, A for 10 subjects using $\xi=0{\cdot}55$, B for 10 subjects using $\xi=0{\cdot}65$ and C for 10 subjects using $\xi=0{\cdot}75$. All of the subjects run in Condition A maintained ρ_0 between 0·5 and 0·6; those in Condition B maintained ρ_0 between 0·6 and 0·7; and those in Condition C maintained ρ_0 between 0·7 and 0·8. The means of the T values were as follows:

	Mean	*S.D.*	*Median*
Condition A ($N=10$)	247·4	38·34	245·5
Condition B ($N=10$)	386·5	68·78	404·5
Condition C ($N=10$)	503·4	79·89	514·5

Applying Jonckheere's (1954) trend test to the original data, $S/\sigma = 4{\cdot}7$ and the null hypothesis is rejected in favour of the predicted trend $T_C > T_B > T_A$ ($P < 0{\cdot}001$).

It thus appears that T is least (and consequently that the learning rate is greatest) when ξ is least (and ρ_0 is least).

Discussion

Taken alone, this result indicates that learning is enhanced when the subject works under high information loading conditions (when he must attend to and deal with problems that are relatively difficult at a given stage in the learning process). Combined with the observations from the pilot experiment, the result suggests that learning is most effective when the subject is *maximally* loaded but *not* overloaded.

This result is consistent with the hypothesis previously advanced. It is also possible, but we believe mistaken, to interpret it as evidence that a subject learns most rapidly in conditions that force him to make a large (but less than critical) number of errors. The fact is that the subject is not really penalized for making an error; many of the events that M counts as ' incorrect ' are ' omissions '; although M dichotomizes events into correct and incorrect responses, the subject observes a ' knowledge of results ' signal containing several possible degrees of correctness. On the other hand, the data do suggest that in these conditions the delivery of knowledge of results is far from a simple reinforcing act and there is plenty of evidence that fairly elaborate features of the situation (such as an increase in η) serve as higher order reinforcers.

Observation of Participant Interaction If a subject accepts the experimental instructions he tacitly agrees to a certain legal interpretation of the events in the experimental situation. He agrees, for example, that stimuli are interpreted in the context of Ω as posing problems and he agrees to construct responses that, so far as possible, denote Ω satisfying solutions to these problems. This body of agreement may be summarized by an assertion that the subject uses an object language L^0 that is concerned with posing and solving problems.

The misuse or illegal use of L^0 in ' participant interaction ' was briefly considered earlier in the paper. It is revealed by the occurrence of repetitive and often idiosyncratic responses that count as mistakes to M but that are not, in a functional sense, mistaken. In retrospect, a subject who behaves in this fashion agrees that he was playing with M; that he responded in a manner designed to make M alter the value of η rather than responding to denote the solution of a problem posed by the stimulus. Further, the idiosyncratic response behaviour is correlated with changes in η (a particular response pattern terminates, for example, when η decreases in value).

It is difficult to quantify ' participant interaction ' in the present situation (though some experiments intended to achieve quantification are in progress).

However, it has been observed: (1) that when 'participant interaction' occurs, it is inhibited if F^0 is replaced by a probabalistic rule (in which the *probability* that $\Delta\eta(n) = +1$ or -1 depends upon $\rho(n)$; in these conditions the subject cannot ascertain the immediate consequences of adopting a particular response pattern); (2) that the incidence of 'participant interaction' is greater when ξ is high and when, as a result, ρ_0 is high. In these conditions the subject has an opportunity to play with the system since he is not so much preoccupied with the immediate matter of responding.

Thus, returning to the issue of stability, it can be argued that systems characterized by a high value of ρ_0 are *less* stable than those with a low value of ρ_0, in so far as they are more liable to the effects of 'participant interaction'.

3. EXPERIMENT 2

Let us call the skill of applying a transformation rule such as Ω a 'sub-skill'. Experiment 2 was concerned with the instruction of complex skills built up from more than one sub-skill. As he learns to perform the task, a subject first changes his attention (to deal with each sub-skill in turn) and later builds up a 'conceptual organization' in which the sub-skills are fully integrated.

The simplest exemplar of a complex (code learning) skill is generated by choosing a pair of transformation rules, Ω_1 and Ω_2 and requiring the subject to apply them alternately. The rules determine a pair of sub-skills, 1 and 2.

At the *n*th trial, the subject is presented with (1) a stimulus and (2) the name of the rule he is to apply. The terminal criterion requires the subject to deal successfully with a situation in which (i) each stimulus is unsimplified and (ii) the rules or sub-skills are rehearsed equiprobably but in such a way that the subject is unable to predict which of them will be needed at a given trial.

The transformation rules were chosen so that certain parts were shared (hence, so that some positive transfer of training would be likely to occur between the sub-skills). Specifically, Ω_1 is the permutation

$$a, 2;\ b, 4;\ c, 1;\ d, 3;\ e, 6;\ f, 8;\ g, 5;\ h, 7;$$

and Ω_2 is the permutation

$$a, 2;\ b, 4;\ c, 1;\ d, 3;\ e, 7;\ f, 5;\ g, 8;\ h, 6.$$

Even though the first part (four members) of each rule is the same, interference between the respective sub-skills dominates positive transfer of training effects in the conditions of the experiment.

As in Fig. 5, the M controlling the complex skill includes duplicates, M^0_i of the control mechanism described for Experiment 1 (one mechanism, M^0_1 in M, is responsible for sub-skill 1 and the other, M^0_2, for sub-skill 2). Thus, if ρ_1, ρ_2 and η_1, η_2 are the proficiencies and difficulties associated with sub-skills 1 and 2, then M^0_1 maintains the condition $\rho_1(n) \rightarrow \rho_0$ and M^0_2 maintains $\rho_2(n) \rightarrow \rho_0$ by an adjustment of η_1 and η_2. In addition, M is equipped with a higher level mechanism, M^1, containing a control rule, F^1, which alternates the sub-skills (and, at a given trial, selects either M^0_1 or M^0_2 as the relevant control system).

Experiment 2(*i*)

In order to compare the acquisition of a code alternation skill in two different conditions, A and B (described below), the stimulus sequence was divided into six trial blocks separated by rest intervals of 5 sec. each. Before the beginning of a block, the subject was informed whether Ω_1 or Ω_2 was to be applied, and this instruction persisted until the next announcement (at the beginning of the next block). Although this arrangement reduces the amount of interference between sub-skills 1 and 2 below the level assumed if the transformation rule is liable to be changed at any trial, it still seems impossible for the subject to achieve an adequate measure of proficiency unless he forms a concept that overcomes interference and allows him to regard the job as a whole.

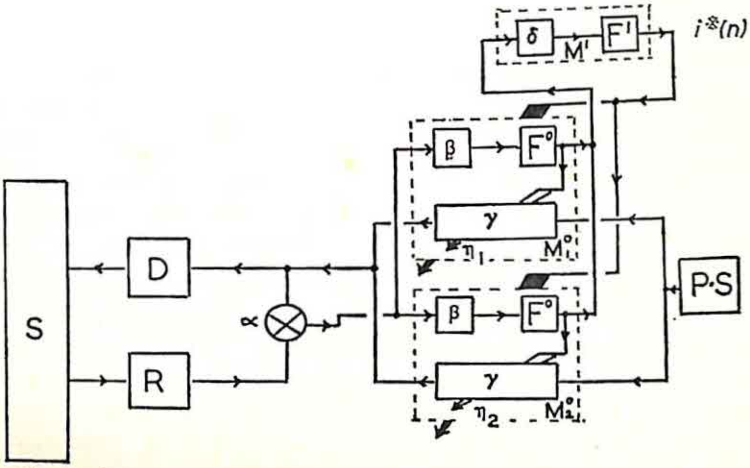


FIGURE 5. Control system used in Condition A for sub-skill alternation. Key as in Fig. 3. δ, computation for alternation rule F^1.

The two conditions, which are juxtaposed in the experiment, differ in the design of M and the type of communication permitted between the subject and M. In Condition A, M is designed according to an hypothesis that rapid and alternating rehearsal of the sub-skills is a prerequisite for the attainment of an overall conceptual organization. Specifically, M detects the subject's proficiency in respect to each sub-skill and rehearses more frequently the sub-skill at which the subject is least proficient. This alternation strategy is embodied in a higher level control rule, F^1. There is, of course, an even less elaborate alternation rule than F^1; namely, the equiprobable rehearsal of the sub-skills for all values of n so that (if $\pi_i(n)$ is the frequency of rehearsal of sub-skill i at the nth trial) $\pi_i(n)=\pi_i=\frac{1}{2}$. This condition has not been included in the present experimental design because it is known, from previous work (Pask, 1964) that $\pi_i(n)=\pi_i=\frac{1}{2}$ yields a slightly but significantly lower rate of learning (that is, a significantly higher value of T) than F^1. In a study that employed similar sub-skills (limited at $m=3$, however, rather than $m=4$) the trend of T values amongst a group of 20 subjects learning by F^1 and a group of 20 subjects learning by equiprobable rehearsal was found to be significant at the 0·5 per cent level (Pask & Lewis, 1966).

In Condition B, the subject is allowed to choose which sub-skill he will rehearse (within the limits that ensure the occasional selection of each sub-skill). Under these circumstances, there is a strong initial tendency for the subject to tackle one job at once (though, ultimately, he settles down to alternation between the sub-skills).

The main predictions were thus: (1) that learning to a common criterion would be appreciably faster in Condition A than in Condition B; and (2) that in Condition A the subject would be prone to receive about the same amount of assistance from M for each sub-skill (this should be an indirect consequence of M's alternation strategy and the separate stabilization of ρ for each sub-skill). In contrast, a Condition B subject (choosing his own alternation strategy) may receive much more assistance in respect of one sub-skill. 'Amount of assistance' can be numerically interpreted (for sub-skills i, up to trial n) as the area beneath the graph of η_i up to the nth trial. The 'difference between the assistance received in respect ot sub-skill 1 and sub-skill 2' is the difference, at trial n, between the areas beneath the graphs of η_1 and η_2. This quantity is designated $\lambda(n)$ and the hypothesis cited above is simply that for $n = T$, the mean of $\lambda(T)$ for Condition A will be significantly less than the mean of $\lambda(T)$ for Condition B.

It was also predicted, though with less confidence, that in Condition A the subject would learn to perform sub-skill 1 in the same way as sub-skill 2, whereas this statement would not usually be true for Condition B. Response latencies were recorded to provide a more detailed picture of the terminal behaviour for each sub-skill in each condition of the experiment. The matter is taken up in the discussion, where it is argued that the terminal latency patterns for sub-skills 1 and 2 are similar in Condition A though they are not in Condition B.

Apparatus

Display and Response Board Modifications The display and response arrangements of Experiment 1 were slightly modified for use in Experiment 2.

The sets of signal lamps, A, B and C and the response buttons, D, were unaltered. An additional lamp, E, situated at the top of the board, was provided to indicate the end of the nth six-trial block. Next, there was a lamp pair, F, situated just below E. For each block, one or the other lamp was illuminated to indicate that Ω_1 or Ω_2 must be applied in the current block of trials. Certain other facilities were optional (being used in Condition B but not in Condition A). These facilities were two indicators, G and H (representing the values of η_1 and η_2), and two auxiliary response buttons, I and J, with which (in the rest interval after a trial block) the subject expressed his preference for rehearsing sub-skill 1 or 2 in the subsequent block of trials.

Apart from the interpolation of a 5-sec rest interval between each six-trial block, the sequencing of events was exactly as it was in the original experiments.

Design of M M is elaborated in several ways, as shown in Fig. 5. In the first place, there is a separate proficiency measure, ρ_1 and ρ_2 for each sub-skill and this is used to determine a separate degree of simplification, μ_1, μ_2 (or of difficulty, η_1, η_2) by the sub-systems $M^0{}_i$, of M. However, the same control rule, F^0, is used for each sub-system, $M^0{}_i$, and ρ_0 is chosen, on the basis of Experiment 1, to minimize the expected value of T. Hence, ρ_0 approaches $\rho_{\min}$ (this choice involves an obviously inaccurate assumption of quasi-independence between the sub-skills).

Next, the sub-skills are alternated. For Condition A this alternation is effected by a higher level control rule, F^1, that is built into M^1 of M in Fig. 5, and which has the form:

$$\begin{aligned} i(n+1) &= \text{not } i(n) \quad \text{if} \quad r(n)=5 \\ &= 1 \quad \text{if} \quad 5>r(n) \quad \text{and} \quad \eta_2(n)>\eta_1(n) \\ &= 2 \quad \text{if} \quad 5>r(n) \quad \text{and} \quad \eta_1(n)>\eta_2(n) \\ &= 1 \quad \text{with probability of } \tfrac{1}{2} \text{ or} \\ &\ 2 \quad \text{with probability of } \tfrac{1}{2} \\ &\qquad\qquad \text{if} \quad 5>r(n) \quad \text{and} \quad \eta_1(n)=\eta_2(n). \end{aligned}$$

Here $i(n)$ names the sub-skill to be rehearsed throughout the nth block of trials; $\eta_1(n)$ and $\eta_2(n)$ are degrees of difficulty associated with sub-skills 1 and 2 by the quasi-independent application of F^0; finally, $r(n)$ is the number of consecutive previous occasions upon which the value of the index i has not been changed. The restrictions entailing $r(n)$ ensure that sub-skills 1 and 2 are sampled regardless of the state of the coupled ' subject, M ' system.

Control rule F^1 is a member of a class of control rules intended to maximize the product of variables such as the $\eta_i(n)$. These rules lead to the most frequent rehearsal of the sub-skill at which the subject is least proficient and if the subject learns the skill, they yield convergence to a terminal condition in which, at trial $n=T$,

$$\eta_1(T)=\eta_2(T)=\eta_{\max}, \quad \rho_1(T)=\rho_2(T)>\rho_0, \quad \pi_1(T)=\pi_2(T)=\tfrac{1}{2}$$

(where $\pi_i(n)$ is the frequency with which the ith sub-skill is rehearsed). The terminal criterion is specified by adding the stipulation that $\rho_1(T)>0{\cdot}75$ and $\rho_2(T)>0{\cdot}75$ to the terminal condition cited above.

Condition B allows the subject, rather than M, to exercise control over the alternation of the sub-skills. The subject was told that he must ultimately satisfy the terminal criterion and he was told, also, that he must rehearse each sub-skill for at least one block of trials in each group of five blocks of trials; the latter instruction constrains his activity in exactly the same way as the r-dependent part of the M control rule F^1. He was provided with the values of $\eta_1(n)$ and $\eta_2(n)$ through the indicators G and H and was further instructed to press either I or J, but not both, in each rest interval when lamp E is illuminated (thus selecting one sub-skill or the other for rehearsal throughout the subsequent trial block). To avoid procedural mistakes or disobedience, a buzzer sounded if neither I nor J was pressed whilst E was illuminated or if both I and J were pressed. This buzzer also sounded if the subject tried to select the same sub-skill on more than four consecutive trial blocks (in fact, if such a selection had been made, it would have been rejected automatically by the system; however, none of the subjects did persist in disobeying the sampling rule). These arrangements are schematized in Fig. 6 wherein the communication channel involving G, H, I and J should be regarded as the medium for a higher level of discourse, L^1.

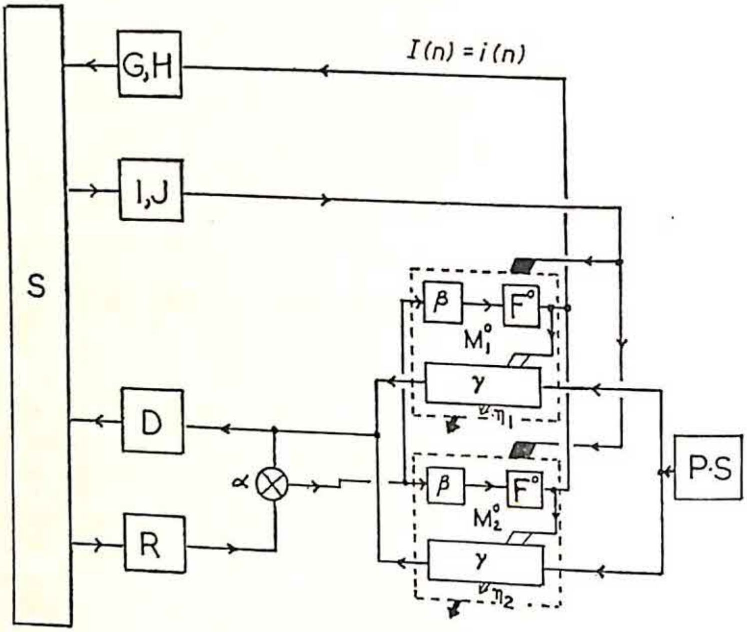


FIGURE 6. Control system used in Condition B for sub-skill alternation. Key as in Fig. 3. G and H, displays of η_1, η_2; I and $\mathcal{J}$, preference assertion responses.

Subjects Twenty subjects were used for Experiment 2(i); 10 of them being assigned to Condition A and 10 to Condition B. The subjects came from the same population as those used in Experiment 1.

Procedure Before the experiment started, all of the subjects were familiar with Ω_1 and Ω_2 and were inveigled into a fairly lengthy discussion about the effects of interference and the need to overcome it in order to achieve the terminal criterion.

The 10 subjects assigned to Condition A were told that M would alternate the sub-skills in a manner designed to ensure overall competence and the most rapid possible approach to the terminal criterion. The 10 subjects assigned to Condition B were told about the various procedural rules, about the displays G and H and about the buzzer. The experimenter stressed the belief that it would be necessary to alternate the sub-skills in order to achieve complete proficiency. (However, the subjects were in no way forced to alternate the sub-skills.)

For each group of subjects, the experiment was continued until the terminal criterion had been achieved and until $\rho_1(n) > 0{\cdot}75$ and $\rho_2(n) > 0{\cdot}75$ for four consecutive blocks of trials.

The values of $x(n)$, $y(n)$, $\eta_1(n)$, $\eta_2(n)$, $\rho_1(n)$, $\rho_2(n)$ and $i(n)$ were recorded electromechanically. For eight subjects it was also possible to record response latencies for each component of the constructed response.

Results

Typical behaviour in Condition A is shown in Fig. 7. One salient difference between Conditions A and B is that T is less (learning rate is greater) in Condition A than it is in Condition B and the generality of this result is exhibited by the following data, comparing the performance of the 10 subject groups run in these conditions. The T means are in terms of six trial blocks.

	Mean of T	*S.D.*	*Median*
Condition A ($N=10$)	76·2	13·14	76
Condition B ($N=10$)	116·4	20·81	109

Using the Mann–Whitney test, the values of T_A and T_B were found to be significantly different ($U=3$; $P<0{\cdot}001$). Hence, Hypothesis (1) (learning rate) is supported by the data.

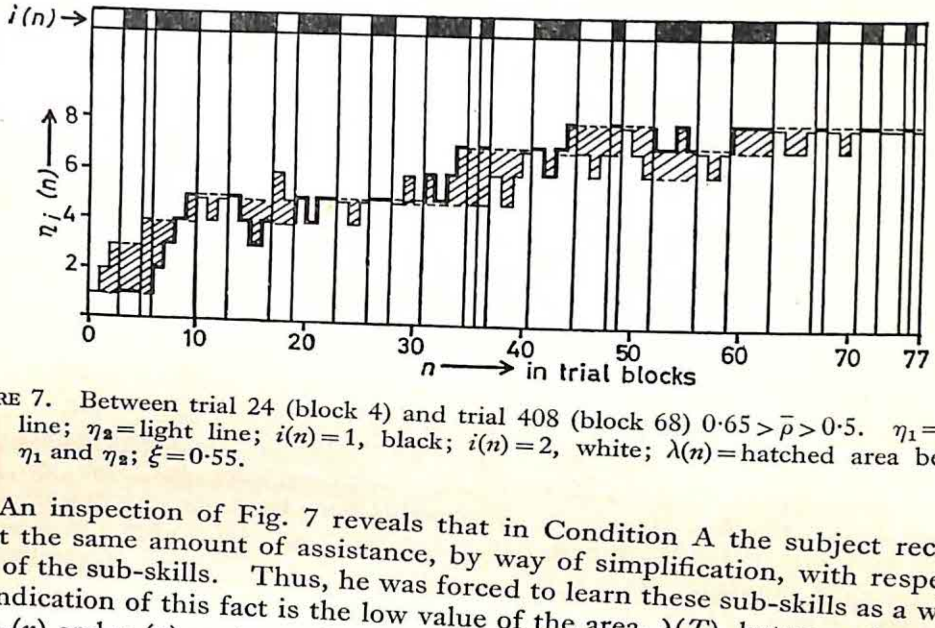


FIGURE 7. Between trial 24 (block 4) and trial 408 (block 68) $0{\cdot}65 > \bar{p} > 0{\cdot}5$. η_1 = heavy line; η_2 = light line; $i(n)=1$, black; $i(n)=2$, white; $\lambda(n)$ = hatched area between η_1 and η_2; $\xi=0{\cdot}55$.

An inspection of Fig. 7 reveals that in Condition A the subject received about the same amount of assistance, by way of simplification, with respect to each of the sub-skills. Thus, he was forced to learn these sub-skills as a whole. An indication of this fact is the low value of the area, $\lambda(T)$, between the curves for $\eta_1(n)$ and $\eta_2(n)$ up to $n=T$. In contrast, subjects run in Condition B were prone to rehearse each sub-skill separately, until a point, usually around block 80, when they found the effort fruitless and adopted an alternation strategy. Hence, for the subjects assigned to Condition B, the value of $\lambda(T)$ is higher than it is in Condition B and the generality of this result is indicated by a comparison of the values of $\lambda(T)$ and $\lambda(T)/T$ for the subjects in Conditions A and B.

	$\lambda(T)$		$\lambda(T)/T$	
	Mean	*S.D.*	*Mean*	*S.D.*
Condition A ($N=10$)	63·4	22·21	0·81	0·16
Condition B ($N=10$)	175·7	53·35	1·53	0·27

The Mann–Whitney test, applied to pairs of $\lambda(T)/T$ values, showed a significant difference between Conditions A and B ($P<0{\cdot}001$), thus supporting Hypothesis (2) concerning the relative amount of assistance given by M.

Latency Patterns To investigate the remaining hypothesis (that in Condition A the subject should learn to perform sub-skill 1 in the same way as sub-skill 2), it was necessary to obtain a more detailed picture of the subject's behaviour in respect to each sub-skill. This was obtained by studying the latencies of the several elements of any response selection. An inspection of the data shows that the response latencies are clustered around discrete values. Where correct responses are concerned, the clustering reflects the pattern in which the m elements of a correct response are selected.

The distribution of response latency patterns has been estimated by dividing Δt into four intervals and counting the response selections that fall within a single interval as 'coincident' selections. Thus, in the case of $m=4$ (namely the values $\eta_i=7$ and $\eta_i=8$), it is possible to distinguish between eight response latency patterns (or, neglecting the sequential ordering of components, the five classes, a_4, b_4, c_4, d_4, e_4, indicated below):

e_4 = quadruple of four 'coincident' responses;
d_4 = 'triple + single' or 'single + triple';
c_4 = 'double + double';
b_4 = 'double + single + single' or 'single + double + single'
or 'single + single + double',
a_4 = 'single + single + single + single'.

Using the same classification, there are three classes (d_3, b_3 and a_3) for $m=2$, two classes (b_2 and a_2) for $m=2$, and one class (a_1) for $m=1$.

For each subject, each sub-skill, i, and each value of η_1, the number of correct response patterns in each class is counted and these numbers are next expressed as a percentage of the total number of relevant correct response patterns. The result is the sequence of histograms (one for a subject in Condition A and one for a subject in Condition B) in Fig. 8. This sequence reflects the acquisition of the skill, the performance of a subject being separately assessed with reference to the rehearsal of sub-skills 1 and 2. (Data are available for four subjects in Condition A and four in B.)

Discussion

Latency Patterns and Similarity between the Terminal Behaviours for Sub-skills 1 and 2 Suppose that different sequences of latency patterns symptomize different methods of learning (e.g., see Pask, 1964, 1965 *a*, *b*). Given this supposition, the latency data indicate that different subjects learn the skill in different ways and that (apart from some ultimate convergence) the subjects learning in Condition A are prone to deal with sub-skills 1 and 2 in a different fashion.

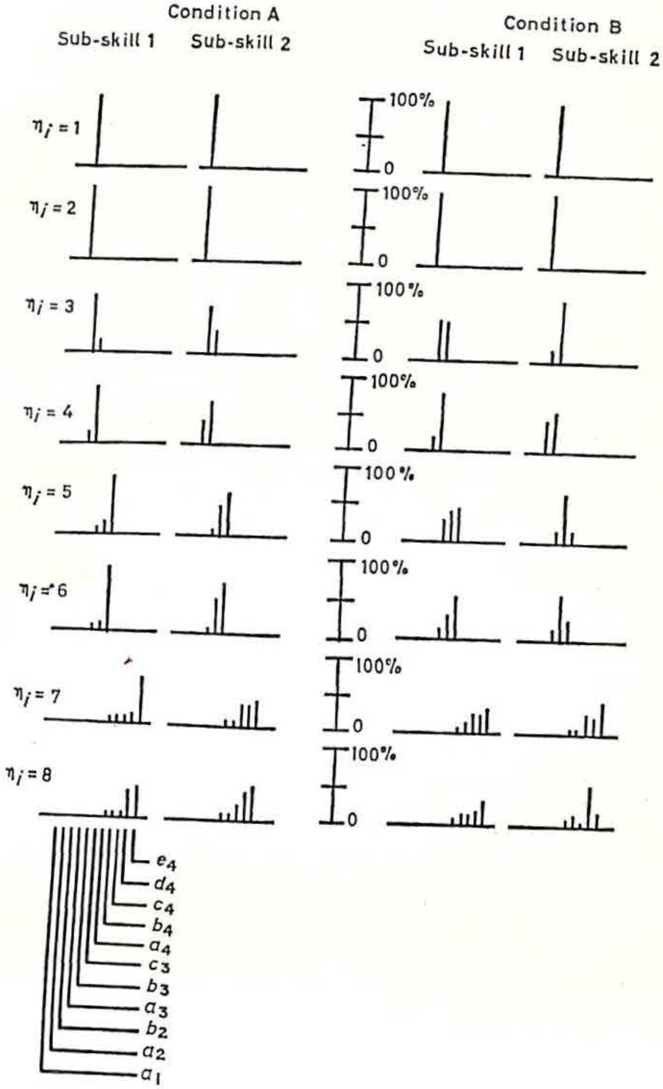


FIGURE 8. Histograms showing the development of latency patterns. One subject in Condition A and one subject in Condition B. Sub-skills 1 and 2 are represented separately.

However, if a subject acquires a common concept that is applicable to sub-skills 1 and 2, his correct response latency pattern for sub-skill 1 may be expected to approach his correct response latency pattern for sub-skill 2. Further, this tendency should be in very much greater evidence in Condition A, where the subject is forced to alternate the sub-skills, than in Condition B, where this constraint is absent. The hypothesis is supported by the following rather crude comparison between the four subjects in Condition A and the four in B. A rank correlation (Spearman's ρ) was obtained between the correct response percentages in sub-skills 1 and 2 for the terminal states ($\eta_i = 7$ and $\eta_i = 8$). For Condition A, the ρ values are high and for Condition B they are lower. The t test applied to the difference between the coefficients of correlation indicates that this difference is modestly significant ($P < 0{\cdot}05$).

Stability of the System In conclusion, it is necessary to make some informal comments about the stability of the 'subject, M' system. The system used in Condition A is stable in respect to arbitrary negative perturbations of either η_1 or η_2, the effect of the perturbation being compensated by a positive excursion of the unperturbed variable (that is, η_1 by η_2 and vice versa, not η by ρ as in Experiment 1). Since the system is stabilized at the lowest possible value of ρ_0, stability in respect to arbitrary positive perturbations of the η_i would be fortuitous.

The system in Condition A is prone to instability engendered by participant interaction (the illegal usage of L^0) because M is elaborate enough to give the subject something to 'play with' (not surprisingly, he is usually *aware* that he is 'playing with' M). As in Experiment 1, 'participant interaction' is encouraged by increasing ρ_0.

In Condition B, participant interaction is 'legalized', for the subject is allowed to communicate with M via the L^1 channel (to make and receive 'meta-statements' or 'statements about the L^0 interaction'). However, as noted already, the subject's unfettered control over the alternation strategy prolongs learning.

Experiment 2(*ii*)

In this experiment, the subject was allowed contingent (rather than unfettered) control over the alternation strategy in an attempt to 'legalize' participant interaction, without prolonging the learning process.

The control system was an 'adaptive meta-system' in which M exerted roughly the same sort of control over the L^1 discourse that it exerted over the L^0 discourse. M is modelled upon a tutor who responds more or less sympathetically to a student's comments about the conduct of a tutorial, on some occasions accepting suggestions that the topic should be changed, on other occasions rejecting them. The simplest M capable of performing this task forms the adaptive meta-system shown in Fig. 9. Logically, this is a hierarchically organized control system, in which there is an L^1 control loop as well as an L^0 control loop.

Psychologically, there is a control system in which the subject is allowed to engage in *legal* 'participant interaction'. The importance of this added facility is apparent when it is noticed that L^1 discourse is precisely the type of interaction whereby the subject constructs the goals that he uses in forming concepts. Hence, the adaptive meta-system is a system that allows us to externalize this crucial feature of concept learning.

In general, $\theta(n)$ may be any suitable function of the conjoint satisfaction of the L^0 goals and, in general, it may be used in any way that gives the subject a degree of L^1 control that is proportional to his ability to satisfy the L^0 goals.

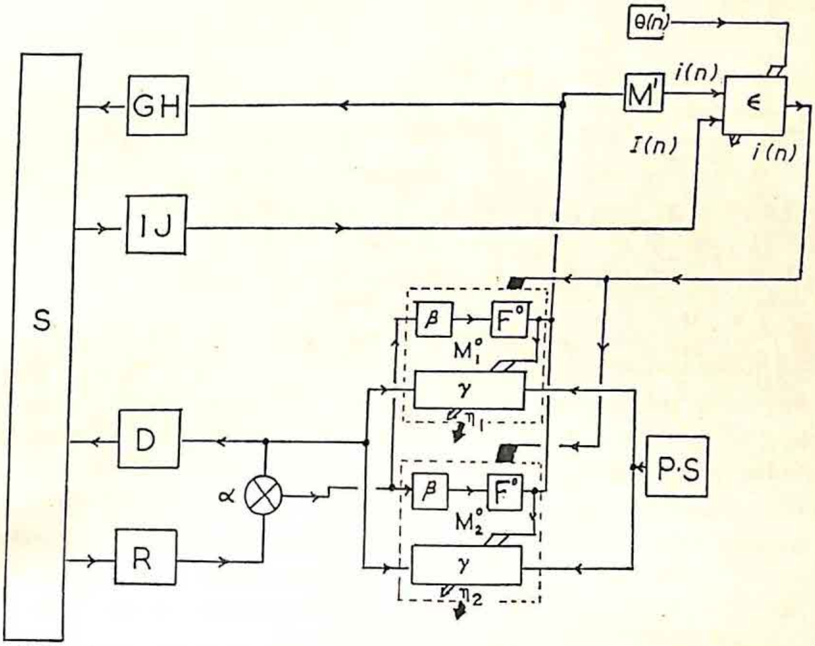


FIGURE 9. Adaptive meta-system. ϵ, electronic dice-thrower biased by $\theta(n)$ for selecting $i(n)=I(n)$ or $i^*(n)$. Other notation as in Figs. 5 and 6.

Method

As in Condition B of Experiment 2(i), the subject made a statement, say $I(n)$, for each block of five trials. $I(n)$ designates one sub-skill or the other chosen for rehearsal and it is interpreted as a preference statement, based upon the L^1 evidence. In addition, M makes a 'preference' statement, say $i^*(n)$, based upon the same evidence. Often, $I(n)$ and $i^*(n)$ are not identical and, in this case, the system in Fig. 9 accepts $I(n)$ with a probability of $\theta(n)$ and $i^*(n)$, with a probability $1-\theta(n)$, the accepted statement which determines the sub-skill for rehearsal being denoted $i(n)$ as before.

The variable $\theta(n)$ has the properties: $\theta(0)=0$, $\theta(T)=1$, $1 \geqslant \theta(n) \geqslant 0$, and $\theta(n)$ increases monotonically as the terminal criterion is approached.

For the present purpose, the form of $\theta(n)$ was chosen as

$$\theta(n)=\rho_1(n)\cdot\rho_2(n)\cdot\eta_1(n)\cdot\eta_2(n)\cdot\left[1-\left|\pi_i(n)-\tfrac{1}{2}\right|\right].$$

Subjects A group of 10 subjects was used for this experiment. They came from the same population as the subjects used in the previous experiments.

Procedure The instructions are identical with those used in Condition A of Experiment 2(i), except for an assertion that the acceptance of the subject's L^1 statements would be contingent upon the value of $\theta(n)$.

Results and Discussion of Experiment 2(*ii*)

Involvement in an adaptive meta-system has a profound motivational effect upon the subject. The learning situation becomes game-like and the game is legal, in contrast to the incidental play of 'participant interaction'. The rate of learning achieved in this experiment (estimated by $1/T$) is greater than the rate achieved in Condition A of Experiment 2(i). This result is not surprising if the 'adaptive meta-system' really does act as it is meant to and, in particular, if it eliminates the unwanted aspects of 'participant interaction'.

For the 10 subjects concerned, the mean value of T is 60 and its S.D. is 9·5 (as compared with the values for Condition A and B of 76·2, S.D. = 13·74 and 116·4, S.D. = 20·81). When the meta-system T values are compared with Condition B (using the Mann–Whitney U test) $U = 0$ ($P < 0{\cdot}001$); the comparison between the meta-system T values and those from Condition A yields $U = 17{\cdot}5$, so that the difference is just significant ($P < 0{\cdot}05$). Jonckheere's (1954) trend test, applied to Conditions A, B, and meta-system, gives $S/\sigma = 4{\cdot}6$ and so the null hypothesis is rejected in favour of the predicted trend

$$T_B > T_A > T_{\text{meta-system}} \ (P < 0{\cdot}001).$$

It is also possible to assert that the form of learning, as reflected in values of $\lambda(T)/T$, conforms closely to the form of learning in Condition A. Once again the conclusion is not at all surprising because M dominates the alternation strategy unless the subject is performing successfully. The mean value of $\lambda(T)/T$ for the 10 meta-system subjects is 0·86, S.D. = 0·28. Using the Mann–Whitney U test, there is no significant difference in respect of $\lambda(T)/T$, between the meta-system and Condition A, but there is a significant difference ($P < 0{\cdot}001$) between the meta-system and Condition B, the inference being that learning in the meta-system is *unlike* the learning in Condition B, but *similar* to the learning in Condition A of Experiment 2(i).

4. Summary and Conclusions

It appears to be possible and useful to study certain types of learning in adaptively stabilized conditions that maintain the value of ρ near to a constant, ρ_0, throughout most of the learning process. In these conditions, the stabilizing procedures that maximize the learning rate are also prone to maximally load the subject (to place him in a state of considerable uncertainty). The same method has been applied to the acquisition of complex skills with several components. Here, learning is enhanced by control strategies that tax the subject heavily and lead him to overcome the effect of interference (between the components of the skill) by acquiring a concept of the skill as a whole. Several regularities in the learning process have been exhibited by using this experimental method. In particular, it is possible to predict the way in which a subject will accept assistance through problem simplification.

Experiment 1 provided data about the acquisition of a simple coding skill, when human learning is compensated by a null-point control mechanism, M.

In Experiment 2(i) a similar technique was used to study the acquisition of a complex skill in which a pair of sub-skills are alternately rehearsed. The comparison of Conditions A and B yielded information about the interaction between the sub-skills and the development of a human conceptual organization. The learning situations involved in these experiments have also been used to illustrate the stability of a coupled ' subject, *M* ' system and to demonstrate an interesting (and apparently ubiquitous) mode of man–machine interaction, ' participant interaction ', which is believed to reflect essential facets of goal construction and concept learning. Experiment 2(ii) introduced one special case of a construction, the ' adaptive meta-system ' which allows the experimenter to investigate ' participant interaction ' as a ' legal activity ' within the experimental framework (rather than regarding it as a mere misuse of the stimulus response modality or object language L^0).

In each experiment, *M* has been viewed as a measuring and stabilizing device which provides otherwise inaccessible data. *M* may also be regarded as a teaching device; its functions in one capacity or the other are logically inseparable. Further, *M* may be generalized as the tool of an experimental method (Pask, 1963 *a*, *b*); the system ' subject, *M* ' does *not* learn and *is* amenable to measurement, even though the system ' subject ' does learn. Using an adaptive meta-system, it is also possible to externalize and measure the goal modifying activity of concept learning.

ACKNOWLEDGEMENTS

This research has been sponsored by the Air Force Office of Scientific Research, O.A.R., under Contract AF61(052)–640, with the European Office of Aerospace Research, U.S. Air Force; and by the Aeronautical Systems Division of the Air Force Systems Command, U.S.A.F., through the European Office of the Office of Aerospace Research, under Contract AF61(052)–402.

REFERENCES

HUDSON, E. M. (1963). An adaptive tracking simulator. *I.R.E. Int. Congr. on Human Factors in Electronics*. (Mimeographed report.)

JONCKHEERE, A. R. (1954). A test of significance for the relation between *m* rankings and *k* ranked categories. *Br. J. statist. Psychol.* **7**, 93–100.

KELLY, C. R. (1966). Design applications of adaptive simulators. *M.I.T., N.A.S.A. Conf. on Manual Control*. Cambridge, Mass.: M.I.T.

LEWIS, B. N. & PASK, G. (1965). The theory and practice of adaptive teaching systems. In R. Glaser (ed.), *Teaching Machines and Programmed Learning*, vol. 2, pp. 213–256. Washington: Nat. Educ. Ass.

PASK, G. (1958 *a*). Electronic keyboard teaching machines. *J. educ. Comm.* Reprinted in E. Lumsdaine & R. Glaser (eds.), *Teaching Machines and Programmed Learning* (1960), vol. 1, pp. 336–348. Washington: Nat. Educ. Ass.

PASK, G. (1958 *b*). Teaching machines. *Proc. 2nd Congr. Int. Ass. Cyber., Namur* pp. 961–978. Paris: Gauthier Villars.

PASK, G. (1960). The teaching machine as a control mechanism. *Trans. Soc. Instr. Tech.*, June, pp. 72–89.

PASK, G. (1961 *a*). *An Approach to Cybernetics*. London: Hutchinson.

PASK, G. (1961 *b*). The simulation of learning and decision making procedure. In C. Muses (ed.), *Aspects of Artificial Intelligence*, pp. 165–210. New York: Plenum Press.

PASK, G. (1963 *a*). A proposed experimental method for the behavioural sciences. In P. Schade & N. Weiner (eds.), *Progress in Biocybernetics*, vol. 1, pp. 171–180. Amsterdam: Elsevier.

PASK, G. (1963 *b*). Tests for some features of a cybernetic model of learning. *Z. f. Psychol.* **171,** 158–195.

PASK, G. (1964). An adaptive teaching system. In K. Austwick (ed.), *Teaching Machines and Programming*, pp. 79–112. London: Pergamon.

PASK, G. (1965 *a*). Teaching as a control engineering process. *Control* (Jan., Feb., March and April).

PASK, G. (1965 *b*). Comments on the organization of man, machines and concepts. In Heilprin, Markusson & Goodman (eds.), *Education for Information Science*, pp. 133–154. New York: Spartan Press; London: Macmillan.

PASK, G. (1966). Man as a system that needs to learn. In D. Stewart (ed.), *Automation Theory and Learning Systems*. London: Academic Press.

PASK, G. & LEWIS, B. N. (1966). The adaptively controlled instruction of a transformation skill. *Programmed Learning*, April 1967, pp. 74–86.

PRINGHORN, M. D. (1961). Perception des distances et constancie des grandeurs. (Thèse 1329, Faculté des sciences, Paris.)

VERHAVE, T. (1963). Toward an empirical calculus for reinforcement value. *J. exp. Analysis Behav.* **6,** 525–536.

WOODWORTH, R. S. (1950). *Experimental Psychology*. London: Methuen.

Vol. 21
Part 1
85–96

The British Journal of Mathematical and
Statistical Psychology

May
1968

NEW METHODS IN MAXIMUM LIKELIHOOD FACTOR ANALYSIS

By K. G. Jöreskog
Educational Testing Service, Princeton
and D. N. Lawley
Department of Statistics, University of Edinburgh

Until recently the main difficulty in the use of maximum-likelihood estimation in factor analysis has been the lack of satisfactory methods of obtaining numerical solutions. This defect has now been remedied, and this paper describes new rapid methods of finding maximum-likelihood estimates.

1. Introduction

In the field of psychology, factor analysis is most often employed to study the measurements that arise from the use of a battery of tests. It will be convenient to discuss factor analysis with particular reference to this type of data, though most of the remarks in this paper are relevant to a much wider context. First of all, a distinction is made between exploratory and confirmatory factor analysis. In confirmatory analysis the experimenter has already obtained a certain amount of knowledge about the variates measured and is therefore in a position to formulate a hypothesis that specifies the factors on which the variates depend. Factor analysis may then be used to test this hypothesis. In exploratory analysis, on the other hand, no such knowledge is available, and the main object is to find a simple but meaningful interpretation of the experimental results. An exploratory analysis is usually performed in two steps. The first step is to decide how many factors are needed to account adequately for the data and to estimate the loadings on the factors, which are initially defined in a somewhat arbitrary manner. A second step consists of a rotation or a linear transformation of these factors into others which can be given a more meaningful interpretation.

In practice, the above distinction is not always clear-cut. Many investigations are to some extent both exploratory and confirmatory, since they involve some variates of known and other variates of unknown factorial composition. The former should be chosen with great care in order that as much information as possible about the latter may be extracted. It is highly desirable that a hypothesis that has been suggested by mainly exploratory procedures should subsequently be confirmed, or disproved, by obtaining new data and subjecting these to more rigorous statistical techniques.

The aim of this expository paper is to give a brief review of two recently developed methods employing maximum-likelihood estimation, one suitable for each type of analysis. Numerical examples illustrating these methods are provided in the last section of the paper. Most of the algebraical details of the underlying theory have been omitted since they can be found in other works referred to in the text.

2. PRELIMINARY CONSIDERATIONS

The basic model in factor analysis is

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{e}, \tag{1}$$

where $\mathbf{x}$ is a column vector of p variates, $\mathbf{f}$ is a vector of k common factors, $\mathbf{e}$ is a vector of p residuals, which represent the combined effect of specific factors and random error, and $\mathbf{\Lambda} = [\lambda_{ir}]$ is a $p \times k$ matrix of factor loadings.

The residuals $\mathbf{e}$ are assumed to be independent of each other and of the common factors $\mathbf{f}$. It is also assumed that the elements of $\mathbf{f}$, $\mathbf{e}$ and $\mathbf{x}$ are all normally distributed with zero means. The dispersion or covariance matrices of $\mathbf{f}$, $\mathbf{e}$ and $\mathbf{x}$ are denoted respectively by $\mathbf{\Phi}$, $\mathbf{\Psi}$ and $\mathbf{\Sigma}$. The matrix $\mathbf{\Psi}$ is diagonal with elements ψ_{ii} $(i = 1, \ldots, p)$, which are termed either residual or unique variances. It is further assumed, without loss of generality, that the common factors have unit variances, so that the diagonal elements of $\mathbf{\Phi}$ are unities. If, in addition, for $k > 1$, the common factors are orthogonal or uncorrelated, then the non-diagonal elements of $\mathbf{\Phi}$ are zeros and thus $\mathbf{\Phi}$ becomes the unit matrix of order k. In view of eqn. (1) and of the assumptions that have been made, $\mathbf{\Sigma}$ is given in terms of the other matrices by the equation

$$\mathbf{\Sigma} = \mathbf{\Lambda \Phi \Lambda}' + \mathbf{\Psi}. \tag{2}$$

This relationship can be tested statistically, unlike eqn. (1), which cannot be verified directly.

Suppose that a random sample of $n+1$ sets of observations of $\mathbf{x}$ is obtained and that $\mathbf{S}$ is the matrix whose elements are the usual unbiased sample estimates of the elements of $\mathbf{\Sigma}$. In view of the assumptions of normality, the elements of $\mathbf{S}$ follow a Wishart distribution with n degrees of freedom. This means that the log-likelihood function L corresponding to the information provided by $\mathbf{S}$ is, neglecting a function of the observations, given by

$$L = -\tfrac{1}{2} n \{\log_e |\mathbf{\Sigma}| + \mathrm{tr}(\mathbf{S\Sigma}^{-1})\}.$$

In order to obtain efficient estimates (for large n) of all unknown parameters L is maximized with respect to these parameters. In practice, it is slightly more convenient to minimize the function

$$F(\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Psi}) = \log_e |\mathbf{\Sigma}| + \mathrm{tr}(\mathbf{S\Sigma}^{-1}) - \log_e |\mathbf{S}| - p. \tag{3}$$

Minimizing F is clearly equivalent to maximizing L, and the minimum value of F multiplied by a constant is later used as a 'goodness of fit' χ^2 criterion.

When $k > 1$, and there is more than one common factor, it is necessary to remove an element of indeterminacy in the basic model before the procedure for minimizing F can be applied. This indeterminacy arises from the fact that a non-singular linear transformation of the common factors changes $\mathbf{\Lambda}$, and in general also $\mathbf{\Phi}$, but leaves $\mathbf{\Sigma}$, and therefore also the function F, unaltered. Hence, in order to obtain a unique set of parameters and a corresponding unique set of estimates, some additional restrictions must be imposed. These have the effect of selecting a particular set of factors and thus of defining the parameters uniquely.

3. EXPLORATORY FACTOR ANALYSIS

Suppose that an entirely exploratory factor analysis is to be performed. Since the correct value for k, the number of common factors, is unknown, it must be determined by a process of trial and error. This will be discussed later. For the present, assume that the value of k is specified and consider the estimation procedure.

The factors **f** will be taken to be orthogonal. Thus $\mathbf{\Phi} = \mathbf{I}$, and $F = F(\mathbf{\Lambda}, \mathbf{\Psi})$ is a function of the elements of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. To define $\mathbf{\Lambda}$ the condition that $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ is a diagonal matrix is imposed. It is supposed also that the diagonal elements of this matrix are distinct and that they are arranged in order of magnitude. This is usually justifiable in practice. Subject to the above condition, the function F is minimized with respect to the elements of $\mathbf{\Lambda}$ and with respect to the ψ_{ii}. By hypothesis, $\mathbf{\Sigma}$ is now given by the equation

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}. \tag{4}$$

The number of parameters in $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ is $pk + p$; but the condition imposed introduces $\frac{1}{2}k(k-1)$ constraints upon them, so that the number of 'free' parameters is

$$p(k+1) - \tfrac{1}{2}k(k-1).$$

For the hypothesis represented by eqn. (4) to be non-trivial this number must be less than $\frac{1}{2}p(p+1)$, the number of distinct elements in $\mathbf{\Sigma}$. This is equivalent to the inequality

$$(p-k)^2 < p + k. \tag{5}$$

The problem of minimizing F (or of maximizing the likelihood) was first considered by Lawley (1940). Since then various other writers have contributed to the subject. A more recent statement of the problem has been given by Lawley & Maxwell (1963). A number of methods have been proposed for maximizing the likelihood. These have all been based upon a direct numerical solution of the equations

$$\partial F / \partial \lambda_{ir} = 0, \qquad \partial F / \partial \psi_{ii} = 0$$

for $i = 1, \ldots, p$ and $r = 1, \ldots, k$. Since these equations cannot be solved algebraically, some iterative procedure has to be used. Unfortunately, most of those previously suggested have not been entirely satisfactory. Convergence to the final solution has often been extremely slow, and in certain cases the procedure has failed to converge. It has been customary to stop iteration when the maximum correction to any parameter is less than a certain value. Such a stopping rule is, however, completely unreliable. As Jöreskog (1966 *b*) has shown, it has on occasions produced supposed solutions that are very inaccurate. Recently a completely new method has been developed by the authors, and this has proved to be completely successful. The full technical details of the method are given in Jöreskog (1966 *b*), and a computer program written by him in FORTRAN IV is also available (1967 *a*). The method concentrates directly on the function F itself rather than on solving the above equations. Though iterative, it converges extremely rapidly, and the values of the parameters which minimize F can be determined as accurately as desired.

Suppose that for given $\boldsymbol{\Psi}$ the function F has a minimum when $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_{\Psi}$, and define the function $f(\boldsymbol{\Psi})$ by

$$f(\boldsymbol{\Psi}) = \min_{\boldsymbol{\Lambda}} F(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = F(\boldsymbol{\Lambda}_{\Psi}, \boldsymbol{\Psi}).$$

Then

$$\min_{\boldsymbol{\Psi}} f(\boldsymbol{\Psi}) = \min_{\boldsymbol{\Lambda}, \boldsymbol{\Psi}} F(\boldsymbol{\Lambda}, \boldsymbol{\Psi}).$$

Thus the problem of minimizing the function F with respect to $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ has been transformed into that of minimizing the function f with respect to the p variables ψ_{ii}.

For given $\boldsymbol{\Psi}$ the numerical determination of $\boldsymbol{\Lambda}_{\Psi}$, using a computer, presents no difficulties. It consists mainly in finding the k largest latent roots and the corresponding latent vectors of the matrix $\boldsymbol{\Psi}^{-\frac{1}{2}}\mathbf{S}\boldsymbol{\Psi}^{-\frac{1}{2}}$. We assume, as is usually the case, that these roots are distinct and greater than unity. The columns of $\boldsymbol{\Lambda}_{\Psi}$ are very simply related to the latent vectors.

The minimization of f is accomplished by using a method of Fletcher & Powell (1963). No attempt will be made to describe this in detail; the essence of the method is that in each iteration a second-degree approximation to the function f is used to estimate the minimum point. This results in a sequence of matrices $\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}, \ldots$ such that

$$f(\boldsymbol{\Psi}^{(s+1)}) < f(\boldsymbol{\Psi}^{(s)}).$$

The sequence converges rapidly to a final matrix of estimates $\boldsymbol{\Psi}$. With each new $\boldsymbol{\Psi}^{(s)}$ there is an associated new $\boldsymbol{\Lambda}^{(s)}$. Thus there is also a sequence of $\boldsymbol{\Lambda}$ matrices, which converges to a final matrix of estimates $\boldsymbol{\Lambda}$.

The procedure begins with an initial approximation $\boldsymbol{\Psi}^{(1)}$ for $\boldsymbol{\Psi}$. It has been shown by Jöreskog that a reasonable approximation is obtained by taking

$$\psi_{ii}^{(1)} = (1 - \tfrac{1}{2}k/p)(1/s^{ii}),$$

where s^{ii} is the ith diagonal element of $\mathbf{S}^{-1}$.

In each iteration the method of Fletcher & Powell requires the calculation of the function value and also the partial derivatives $\partial f/\partial \psi_{ii}$. The latter are easily found, since they are in fact the diagonal elements of the matrix

$$\Psi^{-1}(\Lambda_{\Psi}\Lambda_{\Psi}' + \Psi - S)\Psi^{-1}.$$

The value of f is computed as a function of the latent roots of $\Psi^{-\frac{1}{2}}\mathbf{S}\Psi^{-\frac{1}{2}}$. In addition, the calculation of a positive definite symmetric matrix $\mathbf{E}$ of order p is required. As the iterative procedure converges, the sequence of $\mathbf{E}$ matrices converges to the inverse of the matrix of second-order partial derivatives $\partial^2 f/\partial\psi_{ii}\partial\psi_{jj}$, evaluated at the minimum.

The number of iterations required is considerably reduced by the provision of a good initial estimate of $\mathbf{E}$. A method of obtaining this has been given by Lawley (1967). By standard estimation theory, the final matrix $\mathbf{E}$ multiplied by $2/n$ provides estimates of the sampling variances and covariances of the estimates $\hat{\psi}_{ii}$. In the same paper Lawley has shown how the variances and covariances of the elements of $\hat{\Lambda}$ may also be found.

Various stopping rules for the above procedure could be adopted. In practice, it seems best to stop when the value of each of the first-order partial derivatives is less than a small prescribed value.

When the maximum-likelihood estimates of Λ and Ψ have been found, and f has been minimized, it is possible to test the hypothesis represented by eqn. (4). The minimum value of f is multiplied by the factor

$$n - (2p+5)/6 - 2k/3,$$

and the result is treated as a χ^2 variate for which the number of degrees of freedom is

$$\tfrac{1}{2}\{(p-k)^2 - (p+k)\}.$$

This number is positive provided that inequality (5) is satisfied. The hypothesis is accepted or rejected according to whether the value of χ^2 is below or above a prescribed significance level. This χ^2 test is valid provided that the value of n is reasonably large. A safe rule is that $n \geqslant 50$.

In the above discussion one important point has not been mentioned. Certain data give rise to what has often been termed a Heywood case. In such a case the function f has a minimum only at some point where one or more of the residual variances are negative. To overcome this difficulty, the function f is considered only within the region R_ϵ, where each $\psi_{ii} \geqslant \epsilon$, for some small positive value ϵ. In practice, for standardized variates, we have taken ϵ to be 0·005. If the smallest value of f within R_ϵ is attained on the boundary, so that at least one of the ψ_{ii} is equal to ϵ, the solution is called improper, since in this case f has not attained a true minimum within R_ϵ.

Suppose that an improper solution is obtained in which m of the ψ_{ii} are equal to ϵ. The hypothesis is then reframed, and it is assumed that these residual variances are really zero. The corresponding m variates are eliminated and the partial dispersion matrix calculated for the remaining $p-m$ variates. The first m factors are defined as the m principal components of the m eliminated variates. (If $m=1$, the first factor is simply the eliminated variate after standardization.) The loadings of the remaining $p-m$ variates on the remaining $k-m$ common factors are obtained by an analysis of the partial dispersion matrix of order $p-m$. This yields a proper solution and a $(p-m)\times(k-m)$ loading matrix. The usual χ^2 test is now performed with p, k and n replaced respectively by $p-m$, $k-m$ and $n-m$. The proper solution is finally combined with the information provided by the m eliminated variates to yield a $p\times k$ matrix $\hat{\Lambda}$ which gives maximum-likelihood estimates of the loadings of all p variates on all k factors. The final matrix $\hat{\Psi}$ has zero diagonal elements corresponding to the m eliminated variates. The above process is equivalent to an application of the maximum-likelihood method in which m residual variances are by hypothesis zero.

The frequency with which improper solutions appear to occur in practice is rather surprising. In a study conducted by Mattsson, Olsson & Rosén (1966), some of the results of which are summarized by Jöreskog (1966 *b*), eleven sets of data were analysed by the method described here. In all but two sets, improper solutions were obtained. In some cases these improper solutions had not been suspected when the data were originally analysed, owing to the fact that the iterative procedure then used was stopped too soon.

Since the number of common factors required is initially unknown, a sequential procedure for determining k is used. Starting with some small value k_1, which is often taken to be 1, or even 0, $k=k_1$ is tried. If this value leads to a rejection of the hypothesis, $k=k_1+1$ is next tried, and so on. With each rejection, the value of k is increased by unity and the process continues. An acceptance terminates the process. An upper limit for the value of k is k_2, the largest integer for which inequality (5) is satisfied. If $k=k_2$ is reached and the value of χ^2 is still significant, this means that no non-trivial hypothesis of the form represented by eqn. (4) is acceptable. This sequential procedure is open to certain theoretical objections, but in practice these are unlikely to be very serious and the method seems adequate for an exploratory type of analysis.

It should be emphasized that, in arriving at a final value for k, it has not been proved that this is the true number of common factors present. If the final value of χ^2 is not much above expectation, this implies simply that it would be useless to fit any more common factors, since these would be indistinguishable from experimental error. If, on the other hand, the sequence of tests yields a value of k so large that all factors, after rotation, cannot be given meaningful interpretation, the argument can be advanced that a smaller number of factors should be used even though this does not fit the data. It should also be remembered that, except in artificial sampling experiments, the basic model (1), with its assumptions of linearity and normality, is merely an approximation to reality. The topic of non-linearity has been discussed recently by McDonald (1967). See also other papers referred to therein.

4. CONFIRMATORY FACTOR ANALYSIS

In this section it is assumed that the experimenter has been able to set up a hypothesis that defines the parameters uniquely. The hypothesis must specify the values of certain elements in $\mathbf{\Lambda}$ and in $\mathbf{\Phi}$. It may, in addition, specify the values of some or all of the residual variances ψ_{ii}, though this would be unusual in practice. As a rule, the values specified for the elements of $\mathbf{\Lambda}$ or the non-diagonal elements of $\mathbf{\Phi}$ are zeros, but other values could be used. Let the numbers of specified or fixed parameters in $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ (including diagonal elements of $\mathbf{\Phi}$) be denoted respectively by n_Λ, n_Φ and n_Ψ. Then a necessary, though not sufficient, condition for uniqueness is that

$$n_\Lambda + n_\Phi \geqslant k^2.$$

In general, it is difficult to give sufficient conditions for uniqueness, since the positions of the fixed parameters are important as well as the numbers.

A common type of hypothesis is that certain of the loadings, at least $k-1$ in each column of $\mathbf{\Lambda}$, are zeros and that the diagonal elements of $\mathbf{\Phi}$ are unities; the factors are then correlated or oblique. Another common type of hypothesis is that certain loadings, at least $\frac{1}{2}k(k-1)$ in number, are zero and that $\mathbf{\Phi}$ is the unit matrix of order k, in which case the factors are uncorrelated or orthogonal. It is possible, however, to have hybrid cases in which one group of factors may be correlated while the remaining factors are assumed to be uncorrelated with this group or with each other. The generality of the method gives it great flexibility, since it will deal with all such kinds of hypotheses. Technical details and several examples illustrating the usefulness of the method are given by Jöreskog (1967 *b*).

The total number of parameters in $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ is

$$pk + \tfrac{1}{2}k(k+1) + p = \tfrac{1}{2}(2p+k)(k+1).$$

The number q of fixed parameters is given by

$$q = n_\Lambda + n_\Phi + n_\Psi,$$

and the number of free parameters is

$$\tfrac{1}{2}(2p+k)(k+1) - q.$$

For the hypothesis to be non-trivial this number must be less than $\frac{1}{2}p(p+1)$. This is equivalent to the inequality

$$p^2 + q > \tfrac{1}{2}(p+k)(p+k+1). \tag{6}$$

To apply the method of maximum likelihood, it is necessary to maximize the likelihood, or to minimize the function F, with respect to all free parameters. As before, F is given by eqn. (3), and $\mathbf{\Sigma}$ is given in terms of $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ by eqn. (2). Previous methods for maximizing the likelihood in situations of this kind are referred to by Lawley & Maxwell (1963). In all of these methods, partial derivatives with respect to the free parameters are equated to zero and, after some algebraical simplification, an iterative procedure for solving the equations is employed. Recent work has shown, however, that such procedures do not always converge. Even when convergence does occur it is usually very slow. A better method, for which ultimate convergence is assured, was given by Jöreskog (1966 *a*). Experience with this method has made it clear that it is still sometimes difficult to obtain a very accurate solution unless many iterations are performed. Efficient minimization of the function F seems impossible without the use of second-order derivatives.

The present procedure again uses the method of Fletcher & Powell. Unfortunately, a two-stage minimization procedure such as that described in the previous section is here not possible, except in special cases. The function F has therefore to be minimized simultaneously with respect to all free parameters. The $\mathbf{E}$ matrix evaluated in each iteration converges finally to the inverse of the matrix of second-order derivatives with respect to the free parameters. A method of providing a good initial approximation for $\mathbf{E}$, and thus of reducing the number of iterations required, has been given by Lawley (1967). This involves the calculation and inversion of a symmetric matrix $\mathbf{G}$, whose elements are approximations to the second-order derivatives. In subsequent iterations no further matrix inversion is required, since only simple modifications to $\mathbf{E}$ are necessary.

The order of the matrices $\mathbf{G}$ and $\mathbf{E}$ is the number of free parameters. If this is not too large, the above calculations are easily performed. But if, for example, there were 40 variates and 10 common factors, the number of free parameters might well be almost 400. The inversion and storage of matrices whose order is as large as this present considerable difficulties. With the development of computers having greater storage capacity than those of today, these difficulties may well disappear. It has been found that with a $\mathbf{G}$ matrix of large order a considerable number of non-diagonal elements may reasonably be neglected. This means that a fairly good initial estimate of $\mathbf{E}$ can be obtained by inverting only a number of relatively small sub-matrices of $\mathbf{G}$.

The elements of the final $\mathbf{E}$ matrix multiplied by $2/n$ provide estimates of the sampling variances and covariances of the estimates of the free parameters.

The minimization procedure starts with initial estimates of $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and $\mathbf{\Psi}$. The better these are the fewer iterations will be required. For the most common types of hypotheses good initial estimates are given by the factor transformation methods proposed by Lawley & Maxwell (1964). From the initial point, it is usually best to perform a few steepest descent iterations before employing the method of Fletcher & Powell. Steepest descent iterations have been found to be very effective at the beginning when one is not very close to the minimum. They enable one to obtain better approximations for $\mathbf{G}$ and for the initial $\mathbf{E}$ matrix.

When maximum-likelihood estimates of the free parameters have been found and F has been minimized, it is possible to test the hypothesis represented by eqn. (2) with its specified values for the fixed parameters. The minimum value of F is multiplied by the factor

$$n-(2p+5)/6,$$

and the result is treated as a χ^2 variate for which the number of degrees of freedom is

$$p^2-\tfrac{1}{2}(p+k)(p+k+1)+q.$$

This number is positive provided that inequality (6) is satisfied. The hypothesis is accepted or rejected according to whether the value of χ^2 is below or above the chosen significance level.

The method of this section has been programmed in FORTRAN IV by Jöreskog & Gruvaeus (1967). The program has been tested on an IBM 7044 computer.

5. ANALYSIS OF SOME EMPIRICAL DATA

To illustrate the methods previously discussed, some data of Holzinger & Swineford (1939) are used. The data consist of 26 psychological tests administered to 7th and 8th grade children in two different schools. Only the Grant–White sample is used here. This sample was randomly divided into two samples of sizes 73 and 72 respectively, referred to as the exploration sample and the confirmation sample. A detailed description of the tests is given in the reference above, where miscellaneous descriptive statistics are also given. The following nine tests have been used, the numbers in parentheses being their original code numbers: 1. Visual Perception (1); 2. Cubes (2); 3. Lozenges (4); 4. Paragraph Comprehension (6); 5. Sentence Completion (7); 6. Word Meaning (9); 7. Addition (10); 8. Counting Dots (12); 9. Straight–Curved Capitals (13). The two correlation matrices used in the analysis were computed directly from the test scores.

TABLE 1. TESTS OF HYPOTHESES CONCERNING NUMBER OF FACTORS: EXPLORATION SAMPLE

k	χ^2	*Degrees of freedom*	P
0	253·00	36	< 0·001
1	101·30	27	< 0·001
2	46·39	19	< 0·001
3	5·14	12	0·95

The correlation matrix of the exploration sample was first analysed using the UMLFA program of Jöreskog (1966 *b*). The hypotheses that the number of factors was 0, 1, 2, and 3 were successively tested. The values of χ^2, numbers of degrees of freedom and corresponding probabilities are shown in Table 1. As a general rule, it is probably wise to continue increasing the value of k until the probability P exceeds 0·10. Following this rule, it is concluded that three factors are sufficient to account adequately for the correlations. The maximum-likelihood solution for three factors is given in Table 2. To obtain a preliminary interpretation of the data this solution was rotated orthogonally using the varimax method of Kaiser (1958). The varimax solution is given in Table 3. Since the sample size is rather small, sampling variability is very large. Hence only factor loadings larger than 0·30 in absolute magnitude are interpreted. It then seems that the first factor, determined by tests 1, 2, 3 and 9, is a visual factor, the second factor, determined by tests 4, 5 and 6, is a verbal factor, and that the third factor, determined by tests 7, 8 and 9, is a speed factor.

TABLE 2. UNROTATED MAXIMUM-LIKELIHOOD SOLUTION FOR EXPLORATION SAMPLE

i	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$\hat{\lambda}_{i3}$	$\hat{\psi}_{ii}$
1	0·59	−0·14	0·37	0·49
2	0·37	—0·19	0·45	0·62
3	0·42	−0·32	0·53	0·44
4	0·71	−0·37	−0·27	0·29
5	0·71	−0·26	−0·23	0·37
6	0·74	−0·33	−0·17	0·33
7	0·50	0·58	−0·30	0·32
8	0·65	0·54	0·13	0·27
9	0·64	0·34	0·27	0·40

TABLE 3. VARIMAX ROTATED SOLUTION FOR EXPLORATION SAMPLE

		Factors	
Variate	*Visual*	*Verbal*	*Speed*
1	0·61	0·30	0·23
2	0·60	0·12	0·06
3	0·72	0·18	−0·02
4	0·17	0·82	0·11
5	0·18	0·75	0·21
6	0·26	0·76	0·16
7	−0·22	0·22	0·76
8	0·21	0·12	0·82
9	0·40	0·14	0·65

To examine further the reasonableness of this interpretation, the following target matrix for $\mathbf{\Lambda}$ is set up:

$$\begin{bmatrix} x & 0 & 0 \\ x & 0 & 0 \\ x & 0 & 0 \\ 0 & x & 0 \\ 0 & x & 0 \\ 0 & x & 0 \\ 0 & 0 & x \\ 0 & 0 & x \\ x & 0 & x \end{bmatrix}$$

The matrix of Table 2 was transformed into an oblique solution giving as good agreement as possible with this target. This was accomplished by use of the method of Lawley & Maxwell (1964). The method transforms the factors in such a way that, for any column of the target matrix, the ratio of the sum of squares of loadings corresponding to zeros in the target to the total sum of squares is minimized. The solution is given in Table 4; it is evidently a refinement of that given in Table 3. With a few exceptions the small loadings have become smaller and the large loadings have become larger.

TABLE 4. TRANSFORMED SOLUTION FOR EXPLORATION SAMPLE

		Factors	
Variate	*Visual*	*Verbal*	*Speed*
1	0·60	0·14	0·14
2	0·63	−0·02	−0·00
3	0·74	0·03	−0·11
4	−0·02	0·88	−0·06
5	0·01	0·77	0·05
6	0·10	0·78	−0·00
7	−0·24	0·14	0·78
8	0·26	−0·09	0·82
9	0·44	−0·08	0·63

Factor correlations

	Visual	*Verbal*	*Speed*
Visual	1·00		
Verbal	0·45	1·00	
Speed	0·13	0·36	1·00

TABLE 5. RESTRICTED MAXIMUM-LIKELIHOOD SOLUTION FOR CONFIRMATION SAMPLE

		Factors		
Variate	*Visual*	*Verbal*	*Speed*	*Residual Variance*
1	0·68	0*	0*	0·54
2	0·34	0*	0*	0·88
3	0·66	0*	0*	0·57
4	0*	0·91	0*	0·18
5	0*	0·87	0*	0·25
6	0*	0·82	0*	0·32
7	0*	0*	0·65	0·58
8	0*	0*	0·93	0·15
9	0·67	0*	0·19	0·39

Factor correlations

	Visual	*Verbal*	*Speed*
Visual	1·00*		
Verbal	0·55	1·00*	
Speed	0·47	0·09	1·00*

Asterisks denote parameter values specified by hypothesis.

The results obtained suggest the hypothesis that the nine tests can be explained in terms of three correlated factors with unit variances such that the loading matrix $\mathbf{\Lambda}$ is of the form specified by the target matrix, where 0 now denotes an exact zero loading and x denotes a loading to be estimated from the data. This hypothesis is tested on the confirmation sample, using the RMLFA program of Jöreskog & Gruvaeus (1967). The maximum-likelihood solution under the hypothesis is given in Table 5. The hypothesis is finally accepted, since the value of χ^2 is 29·96 with 23 degrees of freedom, which corresponds to a probability of 0·15.

It should be noted that the solutions of Tables 2–4 are three alternative unrestricted solutions in the same factor space. They fit the observed correlations equally well. The solution of Table 5, on the other hand, is a restricted one. The number of fixed parameters is 20, which is 11 more than that necessary for uniqueness. The restrictions affect the estimation of the residual variances. The differences between the residual variances in Tables 2 and 5 are therefore not entirely due to sampling errors.

The above example has been given mainly to show how the methods described may be put to practical use in cases where the hypothesis is not specified prior to the analysis of the data. If the hypothesis were set up in advance, one could proceed directly to the confirmatory stage.

ACKNOWLEDGEMENT

Part of this work was supported by a grant (NSF–GB 1985) from the National Science Foundation to Educational Testing Service.

REFERENCES

FLETCHER, R. & POWELL, M. J. D. (1963). A rapidly convergent descent method for minimization. *Computer J.* **2**, 163–168.

HOLZINGER, K. J. & SWINEFORD, F. (1939). *A Study in Factor Analysis: The Stability of a Bi-factor Solution.* University of Chicago: Supplementary Educational Monographs, No. 48.

JÖRESKOG, K. G. (1966 *a*). Testing a simple structure hypothesis in factor analysis. *Psychometrika* **31**, 165–178.

JÖRESKOG, K. G. (1966 *b*). UMLFA—A computer program for unrestricted maximum likelihood factor analysis. *Research Memorandum* 66–20. Princeton, N.J.: Educational Testing Service.

JÖRESKOG, K. G. (1967 *a*). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**, 443–482.

JÖRESKOG, K. G. (1967 *b*). A general approach to confirmatory maximum likelihood factor analysis. *Research Bulletin.* Princeton, N.J.: Educational Testing Service.

JÖRESKOG, K. G. & GRUVAEUS, G. (1967). RMLFA—A computer program for restricted maximum likelihood factor analysis. *Research Memorandum* 67–21. Princeton, N.J.: Educational Testing Service.

KAISER, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200.

LAWLEY, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edinb.* (*A*) **60**, 64–82.

LAWLEY, D. N. (1967). Some new results in maximum likelihood factor analysis. *Proc. Roy. Soc. Edinb.* (*A*) **67**, 256–264.

LAWLEY, D. N. & MAXWELL, A. E. (1963). *Factor Analysis as a Statistical Method.* London: Butterworths.

LAWLEY, D. N. & MAXWELL, A. E. (1964). Factor transformation methods. *Br. J. statist. Psychol.* **17**, 97–103.

MCDONALD, R. P. (1967). Factor interaction in nonlinear factor analysis. *Br. J. math. statist. Psychol.* **20**, 205–215.

MATTSSON, A., OLSSON, U. & ROSÉN, M. (1966). The maximum likelihood method in factor analysis with special consideration to the problem of improper solutions. (Research Report, Institute of Statistics, University of Uppsala, Sweden.)

Vol. 21
Part 1
97–103

The British Journal of Mathematical and
Statistical Psychology

May
1968

MOMENTS OF THE STATISTICS KAPPA AND WEIGHTED KAPPA

By B. S. Everitt

Institute of Psychiatry, University of London

This paper considers the mean and variance of the two statistics, kappa and weighted kappa, which are useful in measuring agreement between two raters, in the situation where they independently allocate a sample of subjects to a prearranged set of categories.

1. Problem

In psychological investigations it frequently happens that two or more observers interview the same sample of subjects for the purpose of allocating them to different categories. For example, two psychiatrists may interview a number of patients and independently allocate them to one or other of a number of diagnostic categories. Again, two observers may interview a sample of subjects with regard to a prearranged list of signs or symptoms to determine whether or not the signs are present or absent; and so on.

In such situations, the question arises as to the degree of agreement between the observers. Traditionally, it has been customary to measure agreement by some type of correlation coefficient, such as the ϕ or V coefficient in the case of 2×2 tables (Kendall & Stuart, 1961), or by a coefficient of association or contingency. More recently the problem has attracted renewed attention (Armitage *et al.*, 1966; Cohen, 1960; Fleiss, 1965, 1966).

This paper is concerned with the mean and variance of the coefficient of agreement proposed by Cohen, namely the kappa coefficient. This coefficient is defined as:

$$\kappa = \frac{\text{proportion of observed agreements} - \text{proportion of expected agreements}}{1 - \text{proportion of expected agreements}}.$$

It has also been noted (Spitzer *et al.*, 1967) that κ can be modified to take account of the relative seriousness of the different types of disagreement which can arise between the two observers. This leads to a coefficient called weighted kappa (κ_w) and its mean and variance are also considered.

2. κ AND κ_w

The data appear in the following form:

$$\begin{array}{llcccccc}
 & & \multicolumn{4}{c}{\textit{Diagnostician A}} & & \\
 & & D_1 & D_2 & \dots & D_h & & \\
 & D_1 & m_{11} & m_{12} & \dots & m_{1h} & & n_1 \\
\textit{Diagnostician B} & D_2 & m_{21} & m_{22} & \dots & \dots & & n_2 \\
 & \dots & \dots & \dots & \dots & \dots & & \dots \\
 & D_h & m_{h1} & \dots & \dots & m_{hh} & & n_h \\
 & & M_1 & M_2 & \dots & M_h & & N
\end{array}$$

D_i $(i=1,\dots, h)$ are a set of diagnostic categories,
m_{ij} $(j=1,\dots, h)$ is the observation in the ith row and jth column,
n_i are the row totals,
M_i are the column totals, and the total number of observations is N.

The formulae for κ and κ_w are as follows:

$$\kappa=(P_o-P_c)/(1-P_c), \tag{1}$$

$$\kappa_w=1-\left(\sum_{i=1}^{h}\sum_{j=1}^{h} W_{ij}P_{oij}\Big/\sum_{i=1}^{h}\sum_{j=1}^{h} W_{ij}P_{cij}\right), \tag{2}$$

where P_o = observed proportion of agreement,
P_c = chance expected proportion of agreement,
P_{oij} = observed proportion in cell (i, j) and
P_{cij} = chance proportion in cell (i, j).

In terms of the m_{ij}, n_i and M_i these are:

$$P_o=\frac{1}{N}\sum_{i=1}^{h} m_{ii}, \qquad P_c=\frac{1}{N^2}\sum_{i=1}^{h} h_i m_i,$$

$$P_{oij}=\frac{1}{N}m_{ij}, \qquad P_{cij}=\frac{1}{N^2}n_i m_j.$$

The W_{ij} are a set of weights indicating the level or seriousness of the disagreement. These are arbitrary and chosen by the experimenter. For example, in an experiment to measure the agreement between two psychiatrists in allocating patients to different diagnostic categories, a disagreement such as neurotic disorder versus personality disorder is not as serious as one such as psychotic disorder versus neurotic disorder, and so the first type of disagreement would be allocated a lower weight than the second. (For the diagonal cells the weights are, of course, zero.)

κ may be interpreted as the proportion of agreement over and above that expected by chance. It can be shown to have almost the same value as the product-moment correlation coefficient for the dichotomous case. The advantage of κ appears to be that it is more intuitively reasonable, and also that it leads to κ_w, thus allowing for differences in the seriousness of disagreement.

3. MEAN AND VARIANCE OF κ AND κ_w

The probability density function of the m_{ij} is given by:

$$P(m_{ij}) = \prod_{i=1}^{h} n_i! \prod_{j=1}^{h} M_j! \Big/ \prod_{\substack{i=1\\ j=1}}^{h} m_{ij}!N!. \tag{3}$$

From eqn. (3), the means, variances and covariances of the m_{ij} can be derived by finding the factorial moments of the distribution. However, the covariances may also be found in the following way without assuming any knowledge of factorial moments.

The marginal distribution of a single m_{ij} can be shown to be a hypergeometric distribution given by

$$P(m_{ij}) = \frac{n_i!(N-n_i)!M_j!(N-M_j)!}{N!m_{ij}!(n_i-m_{ij})!(M_j-m_{ij})!(N-n_i-M_j+m_{ij})!}. \tag{4}$$

Therefore the mean and variance of m_{ij} are

$$\frac{n_i M_j}{N} \quad \text{and} \quad \frac{n_i(N-n_i)M_j(N-M_j)}{N^2(N-1)}$$

respectively (Mood & Graybill, 1963, p. 110). Also the sum of any number of the m_{ij} is a hypergeometric variable, and using this fact the covariance of any two of the m_{ij} can be derived. The covariance of m_{11} and m_{12} can be found as follows.

Consider $T = m_{11} + m_{12}$, then

$$\operatorname{var}(T) = \operatorname{var}(m_{11}) + \operatorname{var}(m_{12}) + 2\operatorname{cov}(m_{11}, m_{12}). \tag{5}$$

Now T is a hypergeometric variable with variance given by

$$\operatorname{var}(T) = \frac{n_1(N-n_1)(M_1+M_2)(N-M_1-M_2)}{N^2(N-1)}; \tag{6}$$

and the variances of m_{11} and m_{12} are:

$$\operatorname{var}(m_{11}) = \frac{n_1(N-n_1)M_1(N-M_1)}{N^2(N-1)} \tag{7}$$

and

$$\operatorname{var}(m_{12}) = \frac{n_1(N-n_1)M_2(N-M_2)}{N^2(N-1)}. \tag{8}$$

Therefore using eqns. (6), (7) and (8) it is possible to solve eqn. (5) to give

$$\operatorname{cov}(m_{11}, m_{12}) = \frac{-n_1 M_1 M_2 (N-n_1)}{N^2(N-1)}.$$

Similarly, for any two elements in the same row,

$$\operatorname{cov}(m_{tp}, m_{t,p+s}) = \frac{-n_t M_p M_{p+s}(N-n_t)}{N^2(N-1)} \quad (s \neq 0). \tag{9}$$

In a similar way the covariance for any two elements in the same column can be shown to be

$$\operatorname{cov}(m_{tp}, m_{t+s,p}) = \frac{-n_t n_{t+s} M_p (N-M_p)}{N^2(N-1)} \quad (s \neq 0). \tag{10}$$

For diagonally opposed elements, the same type of procedure is used. For example, to find the covariance of m_{11} and m_{22}, or of m_{12} and m_{21} (these two covariances are equal, as may be seen by considering the probability density function of the m_{ij}), the following procedure can be followed.

Consider $T = m_{11} + m_{12} + m_{21} + m_{22}$. Then

$$\begin{aligned}\operatorname{var}(T) = {} & \operatorname{var}(m_{11}) + \operatorname{var}(m_{12}) + \operatorname{var}(m_{21}) + \operatorname{var}(m_{22}) \\ & + 2\operatorname{cov}(m_{11}, m_{12}) + 2\operatorname{cov}(m_{11}, m_{21}) + 2\operatorname{cov}(m_{11}, m_{22}) \\ & + 2\operatorname{cov}(m_{12}, m_{21}) + 2\operatorname{cov}(m_{12}, m_{22}) + 2\operatorname{cov}(m_{21}, m_{22}).\end{aligned} \tag{11}$$

Again T is a hypergeometric variable with variance given by

$$\operatorname{var}(T) = \frac{(n_1+n_2)(N-n_1-n_2)(M_1+M_2)(N-M_1-M_2)}{N^2(N-1)}. \tag{12}$$

Therefore in eqn. (11) the only unknowns are $\operatorname{cov}(m_{11}, m_{22})$ and $\operatorname{cov}(m_{12}, m_{21})$ and as they are equal they may be found from eqn. (11). They are given by

$$\operatorname{cov}(m_{11}, m_{22}) = \operatorname{cov}(m_{12}, m_{21}) = \frac{n_1 n_2 M_1 M_2}{N^2(N-1)}.$$

Similarly, for any diagonally opposed elements,

$$\operatorname{cov}(m_{st}, m_{s+w,t+u}) = \frac{n_s n_{s+w} M_t M_{t+u}}{N^2(N-1)} \quad (w \neq 0, u \neq 0). \tag{13}$$

Using the results of eqns. (9), (10) and (13), the mean and variance of the statistics κ and κ_w can be derived.

(*a*) *Mean values of* κ *and* κ_w

$$E(\kappa) = \frac{1}{1-P_c}\{E(P_o) - P_c\} \qquad \text{(since } P_c \text{ is fixed by the marginal totals)},$$

where E denotes mathematical expectation.

Now,

$$E(P_o)=\frac{1}{N}\sum_{i=1}^{h}E(m_{ii})=\frac{1}{N}\sum_{i=1}^{h}\frac{n_iM_i}{N}=\frac{1}{N^2}\sum_{i=1}^{h}n_iM_i=P_c.$$

Therefore the expected value of κ is zero.

The expectation of κ_w is given by

$$E(\kappa_w)=1-\left(\sum_{i=1}^{h}\sum_{j=1}^{h}w_{ij}P_{cij}\right)^{-1}E\left(\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}P_{oij}\right);$$

and

$$E\left(\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}P_{oij}\right)=\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}E(P_{oij})$$

$$=\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}\frac{n_iM_j}{N^2}=\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}P_{cij}.$$

Therefore the expected value of κ_w is also zero.

(b) Variances of κ and κ_w

$$\operatorname{var}(\kappa)=\frac{1}{(1-P_c)^2}\operatorname{var}(P_o); \tag{14}$$

$$\operatorname{var}(P_o)=\frac{1}{N^2}\left\{\sum_{i=1}^{h}\operatorname{var}(m_{ii})+2\sum_{i=1}\sum_{j=1}\operatorname{cov}(m_{ii}m_{jj})\right\}. \tag{15}$$

Hence, using the values derived for the variances and covariances of the m_{ij}, the variance of κ may be found.

$$\operatorname{var}(\kappa_w)=\left(\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}P_{cij}\right)^{-2}\operatorname{var}\left(\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}P_{oij}\right), \tag{16}$$

$$\operatorname{var}\left(\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}P_{oij}\right)=\frac{1}{N^2}\left\{\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}^2\operatorname{var}(m_{ij})+2SW_{ij}W_{st}\operatorname{cov}(m_{ij},m_{st})\right\}, \tag{17}$$

where S denotes the appropriate summation over the whole table.

Hence it is possible to calculate the variance of κ_w. The variances given in eqns. (14) and (16) are both rather tedious to calculate, especially that of κ_w, and so easily calculable approximations to these variances are given below.

4. Approximate Variances of κ and κ_w

If it is assumed that P_o and P_{oij} are binomial proportions, then it follows that

$$\operatorname{var}(\kappa)=\frac{1}{(1-P_c)^2}\,\frac{P_o(1-P_o)}{N}, \tag{18}$$

$$\operatorname{var}(\kappa_w)=\left(\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}P_{cij}\right)^{-2}\left\{\sum_{i=1}^{h}\sum_{j=1}^{h}W_{ij}^2P_{oij}(1-P_{oij})\right\}\Big/N. \tag{19}$$

5. NUMERICAL EXAMPLE

		Diagnostician B			
		D_1	D_2	D_3	
Diagnostician A	D_1	106 (0)	10 (9)	4 (5)	120
	D_2	22 (9)	28 (0)	10 (3)	60
	D_3	2 (5)	12 (3)	6 (0)	20
		130	50	20	200

The figures in brackets in the top right-hand corner of each cell are the weights for each cell. For example, the weight attached to a disagreement of the kind D_1 vs. D_2 is 9. For this table and set of weights, $\kappa = 0{\cdot}4286$, $\kappa_w = 5071$.

(*a*) If κ is used,

$$\text{Approximate variance} = \frac{1}{(1-0{\cdot}475)^2}\,\frac{(0{\cdot}700).(0{\cdot}300)}{200} = 0{\cdot}0038;$$

$$\begin{aligned}\text{Exact variance} &= \frac{1}{(1-0{\cdot}475)^2}\,\frac{1}{(200)^2}\,\frac{1}{(200)^2}\,199\\ &\quad\times(120.80.130.70+60.140.50.150+20.180.20.180\\ &\quad+2.120.130.60.50+2.120.130.20.20+2.60.50.20.20\\ &= 0{\cdot}0031.\end{aligned}$$

(*b*) If κ_w is used,

$$\begin{aligned}\text{Approximate variance} &= \frac{1}{(3{\cdot}895)^2}\,\frac{1}{200}\\ &\quad\times\{81(0{\cdot}05)(0{\cdot}95)+25(0{\cdot}02)(0{\cdot}98)+81(0{\cdot}11)(89)\\ &\quad+9(0{\cdot}05)(0{\cdot}95)+25(0{\cdot}01)(0{\cdot}99)+9(0{\cdot}06)(0{\cdot}94)\}\\ &= 0{\cdot}004425.\end{aligned}$$

The exact variance is very tedious to calculate, but it can be shown to be 0·004417. It can be seen that the approximations in this case are very good. However, for smaller values of N, the difference between the approximate and exact variances will be larger.

6. SUMMARY

The means and variances of the statistics, kappa and weighted kappa, used in measuring observer agreement, were found. Easily calculable approximations to the variances were also given, and a numerical example suggests that for a large number of observations these approximations are very good.

ACKNOWLEDGEMENTS

This work was supported by N.I.M.H. Grant No. MH–09191 and was carried out under the U.S.-U.K. Diagnostic Project at the Institute of Psychiatry.

REFERENCES

ARMITAGE, P., BLENDIS, L. M. & SMYLLIE, H. C. (1966). The measurement of observer disagreement in the recording of signs. *J. R. statist. Soc.* (*A*) **129**, 98–109.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educ. psychol. Measur.* **20**, 37–46.

FLEISS, J. L. (1965). Estimating the accuracy of dichotomous judgements. *Psychometrika* **30**, 469–479.

FLEISS, J. L. (1966). Assessing the accuracy of multivariate observations. *J. Am. statist. Ass.* **61**, 403–412.

KENDALL, M. G. & STUART, A. (1961). *The Advanced Theory of Statistics*, vol. 2. London: Griffin.

MOOD, A. M. & GRAYBILL, F. A. (1963). *Introduction to the Theory of Statistics.* New York: McGraw-Hill.

SPITZER, R. L. *et al.* (1967). Kappa and weighted kappa. (To be published.)

Vol. 21
Part 1
105-116

The British Journal of Mathematical and
Statistical Psychology

May
1968

DERIVING COEFFICIENTS OF RELIABILITY AND AGREEMENT FOR RATINGS

By A. E. MAXWELL
Institute of Psychiatry, University of London

and A. E. G. PILLINER
Godfrey Thomson Unit for Educational Research, University of Edinburgh

The paper is concerned with the measurement of internal consistency of rating scales and interviewing schedules, with the assessment of bias between different raters and with coefficients for measuring the degree of agreement between them. Analysis of variance models are first employed, but reference is also made to earlier psychometric techniques and to recent work by Armitage *et al.* and by Fleiss.

1. INTRODUCTION

When a number of psychologists or psychiatrists interview a patient, or group of patients, using a previously agreed schedule of symptoms to determine the presence or absence of these symptoms, or to rate on a previously agreed scale the severity of the symptoms, discrepancies between the results are to be expected. A problem of perennial interest is to assess the nature of these discrepancies and to express the results in a concise way. Problems similar to these arise naturally in many other fields of enquiry—in general medicine (e.g. Fletcher, 1960), in psychological and sociological investigations, in public surveys, and so forth, but they will be discussed here from the psychiatric viewpoint since the problems arose in this context. To avoid cumbersome terminology, the term 'rater' will be used throughout the paper in place of 'psychologist' or 'psychiatrist' or the name of some other expert interviewer, and in all equations 'rater' will be referred to by the letters 'D' or 'd' since the letter 'r' is commonly used to denote a correlation or reliability coefficient.

For clarity of analysis the general problem can be separated into more specific problems depending on the conditions of the investigation. Commonly occurring situations are the following:

I. A number of raters interview the same group of patients to assess the presence or absence or the degree of severity of (*a*) a single symptom; (*b*) a number of symptoms which may not be mutually independent.

II. A number of raters interview a single patient to assess the presence or absence, or the degree of severity of a number of symptoms, which again may not be mutually independent. (This is a special case of I*b*, but it is informative to consider it separately.)

Discrepancies between the results given by raters may vary in a number of ways. For instance, in the case of I*a*, there may be bias in the results in the sense that one rater may systematically record the symptom as being present more frequently or as being more severe than does another rater. There may be interaction between raters and patients in the sense that a rater's biases for individual patients may be larger or smaller than his average bias over the whole group. Though fair agreement between the raters may exist, discernible patterns may be recognizable in the disagreements, i.e. it may not be legitimate to assume that disagreements, or 'errors' are randomly distributed (see Fleiss, 1965). Possible methods of analysing the data to take account of different eventualities are discussed below.

2. CONTINUOUS DATA WHICH CONFORM TO A TWO-WAY ANALYSIS OF VARIANCE MODEL

In situations in which degrees of severity of symptoms are measured on a rating scale (e.g. Hamilton, 1959), or other 'continuous' measuring device, where 'scores' are allocated to each patient, an adequate analysis of the data, at least in Case I*a* above, can generally be achieved by well-known methods of analyses of variance. In this approach, transformations of the data to achieve stability of variance (Bartlett, 1947; Box & Cox, 1964) may often be helpful in meeting the assumptions of the model.

Taking up suggestions made by Jackson (1940), Hoyt (1941) used a two-way analysis of variance on a matrix of scores obtained by f persons on each of g questions. He showed that the correlation between the f totals over the g questions could be estimated by:

$$r=(\bar{P}-\overline{PQ})/\bar{P},$$

in which $\bar{P}$ is the mean square for persons and $\overline{PQ}$ the persons × questions interaction mean square. It is an easy matter to use the Spearman–Brown 'prophecy' formula in reverse and obtain for the average correlation between individual question totals:

$$r'=(\bar{P}-\overline{PQ})/\{\bar{P}+(g-1)\overline{PQ}\}.$$

To illustrate the analysis, it will be assumed that t randomly chosen raters interview a sample of n randomly chosen patients and rate each patient on a scale measuring the severity of a single symptom, X (Case I*a* above with a 'continuous' variable). The results could be set out as in Table 1 where the score X_{ij} in the body of the table is that given to patient i ($i=1, 2, \ldots, n$) by rater j ($j=1, 2, \ldots, t$).

TABLE 1. SCORES ALLOTTED BY t RATERS TO n PATIENTS ON SYMPTOMS X

	Raters						
Patients	1	2 …	j	…	t	*Totals*	*Means*
1	X_{11}	…	X_{1j}	…	X_{1t}	P_1	$\bar{P}_1$
2							
⋮	⋮		⋮		⋮	⋮	⋮
i	X_{i1}	…	X_{ij}	…	X_{it}	P_i	$\bar{P}_i$
⋮	⋮		⋮		⋮	⋮	⋮
n	X_{n1}	…	X_{nj}	…	X_{nt}	P_n	$\bar{P}_n$
Totals	D_1	…	D_j	…	D_t	T	
Means	$\bar{D}_1$	…	$\bar{D}_j$	…	$\bar{D}_t$		$\bar{T}$

A two-way analysis of variance of the data is shown in Table 2. In this analysis the estimated mean squares, denoted by $\bar{D}$, $\bar{P}$ and $\overline{DP}$ are obtained by dividing the sums of squares by their respective degrees of freedom. The final column in Table 2 gives the expected mean squares for the several components of variation: σ_0^2 is the residual or 'error' variance; σ_p^2 is the variance of patients' mean scores, while σ_d^2 is the variance of the mean scores for the raters. On the assumption that the variances of the columns of data in Table 1 are homogeneous, bias between the raters can be tested by the F-ratio $\bar{D}/\overline{DP}$.

TABLE 2. TWO-WAY ANALYSIS OF VARIANCE OF THE DATA IN TABLE 1

Source of variation	*Degrees of freedom*	*Sums of squares*	*Mean square*	*Expected mean square*
Raters	$t-1$	$n\sum_{t}(\bar{D}_j-\bar{T})^2$	$\bar{D}$	$\sigma_0^2+n\sigma_d^2$
Patients	$n-1$	$t\sum_{n}(\bar{P}_i-\bar{T})^2$	$\bar{P}$	$\sigma_0^2+t\sigma_p^2$
Residual	$(t-1)(n-1)$	by difference	$\overline{DP}$	σ_0^2
Total	$nt-1$	$\sum_{t}\sum_{n}(X_{ij}-\bar{T})^2$		

A measure of the overall agreement between this random group of t raters and another random group of t raters is given by the ratio:

$$\rho_1=t\sigma_p^2/(t\sigma_p^2+\sigma_0^2),$$

which can be estimated from the mean squares as

$$r_1=(\bar{P}-\overline{DP})/\bar{P}. \tag{1a}$$

A measure of the average agreement between one rater and another is given by the ratio

$$\rho_2=\sigma_p^2/(\sigma_p^2+\sigma_0^2),$$

which can be estimated from the mean squares as

$$r_2=(\bar{P}-\overline{DP})/\{\bar{P}+(t-1)\overline{DP}\}. \tag{1b}$$

These r's are significant if the F-ratio $\bar{P}/\overline{DP}$ shows $\bar{P}$ to be significantly greater than $\overline{DP}$.

3. CONTINUOUS DATA: THREE-WAY ANALYSIS OF VARIANCE MODEL—CASE I*b*

Consideration is now given to Case I*b* in situations in which rating scales are employed. This case could be dealt with by performing separate analyses, of the type outlined in Section 2 above, for each symptom. But since the symptoms are unlikely to be independent of each other, an analysis of the data which simultaneously takes them all into account is generally more informative. Therefore it will be assumed that t raters ($j=1, 2, \ldots, t$) interview the same n patients ($i=1, 2, \ldots, n$) and rate each on k symptoms ($h=1, 2, \ldots, k$). It will be further assumed that the rating scales for the several symptoms are similar in range and, for the purposes of the model, that all the raters replicate their ratings for each of the k symptoms m times ($q=1, 2, \ldots, m$) in the investigation. The data are set out symbolically, for a single replication, in Table 3.

TABLE 3. SCORES ALLOCATED BY EACH OF t RATERS TO n PATIENTS ON k SYMPTOMS

	First Rater						*t-th Rater*				
Symptoms	1	2 …	h	…	k		1	2 …	h	…	k
Patients											
1	X_{11}	…	X_{1h}	…	X_{1k}	…	Y_{11}	…	Y_{1h}	…	Y_{1k}
2											
⋮			⋮			…			⋮		
i			X_{ih}			…			Y_{ih}		
⋮			⋮			…			⋮		
n						…					
Means	$\overline{X}_{\cdot 1}$	…	$\overline{X}_{\cdot h}$	…	$\overline{X}_{\cdot k}$	…	$\overline{Y}_{\cdot 1}$	…	$\overline{Y}_{\cdot h}$	…	$\overline{Y}_{\cdot k}$

A three-way analysis of variance model with replication is now considered. It is assumed that the score of patient i on replication q of symptom h for rater j can be represented as:

$$x_{ihjq}=\alpha+\beta_i+\gamma_h+\delta_j+\beta\gamma_{ih}+\beta\delta_{ij}+\gamma\delta_{hj}+\beta\gamma\delta_{ihj}+e_{ihjq}. \tag{2}$$

In this model, α is the general mean; β_i the contribution specific to patient i; γ_h the contribution specific to symptom h; δ_j the contribution specific to rater j; the next four terms in eqn. (2) are the interaction terms, while e_{ihjq} is replication error. One purpose of the analysis would be to see if there is bias between the mean scores allocated by raters, to see if some symptoms are scored more highly than others, or if there are significant interactions, etc. Another purpose would be to derive from the analysis coefficients of agreement among raters.

For the purposes of the analysis the most general case will be considered in which raters, symptoms and patients (referred to in the analysis by the letters D, S, and P, respectively) are all considered to be random samples from parent populations. The analysis of variance of the data would now take the form shown in Table 4. For reasons of space, the 'sums of squares' column is omitted, but these would be calculated in the usual way (e.g. Winer, 1962, p. 171) or would more probably be obtained on a computer. In the final column of the table, σ_β^2 stands for the variance between patients; σ_γ^2 for variance between symptoms; σ_δ^2 for variance between raters. Variances due to interactions are indicated by σ^2 with appropriate subscripts, while σ_0^2 stands for the 'error' variance. After making appropriate assumptions about which are 'random' effects and which are 'fixed', the expected mean squares indicate the appropriate error terms to use for testing main effects and interaction effects.

TABLE 4. A THREE-WAY FACTOR ANALYSIS WITH REPLICATIONS

Source of variation	*Degrees of freedom*	*Mean squares*	*Expected mean squares*
Patients (P)	$(n-1)$	$\bar{P}$	$\sigma_0^2+m\sigma_{\beta\gamma\delta}^2+mt\sigma_{\beta\gamma}^2+mk\sigma_{\beta\delta}^2+mkt\sigma_\beta^2$
Symptoms (S)	$(k-1)$	$\bar{S}$	$\sigma_0^2+m\sigma_{\beta\gamma\delta}^2+mt\sigma_{\beta\gamma}^2+mn\sigma_{\gamma\delta}^2+mtn\sigma_\gamma^2$
Raters (D)	$(t-1)$	$\bar{D}$	$\sigma_0^2+m\sigma_{\beta\gamma\delta}^2+mk\sigma_{\beta\delta}^2+mn\sigma_{\gamma\delta}^2+mkn\sigma_\delta^2$
DS	$(t-1)(k-1)$	$\overline{DS}$	$\sigma_0^2+m\sigma_{\beta\gamma\delta}^2+mn\sigma_{\gamma\delta}^2$
DP	$(t-1)(n-1)$	$\overline{DP}$	$\sigma_0^2+m\sigma_{\beta\gamma\delta}^2+mk\sigma_{\beta\delta}^2$
SP	$(k-1)(n-1)$	$\overline{SP}$	$\sigma_0^2+m\sigma_{\beta\gamma\delta}^2+mt\sigma_{\beta\gamma}^2$
DSP	$(t-1)(k-1)(n-1)$	$\overline{DSP}$	$\sigma_0^2+m\sigma_{\beta\gamma\delta}^2$
Replications	$tkn(m-1)$	$\bar{E}$	σ_0^2

In a recent paper Fleiss (1966) considers a multivariate approach to the problem discussed above and provides a likelihood ratio test for assessing bias amongst raters. A criterion is also given for deriving measures of agreement between raters. A distinct advantage of the multivariate approach is that it takes into account possible correlation between the variables.

4. COEFFICIENTS OF AGREEMENT DERIVED FROM THREE-WAY ANALYSIS OF VARIANCE

The use of information provided by the analysis of variance for deriving coefficients of agreement between raters is less well known than its use for comparing means and this will be given in greater detail. Without loss of generality, the mean of the data can be shifted to the origin, so that $\alpha=0$. From eqn. (2)

the total score for patient i over k symptoms each replicated m times and rated by t raters is obtained by summing over j, h and q. It is:

$$X_i = \sum_j \sum_h \sum_q x_{ijhq} = ktm\beta_i + tm \sum_h \gamma_h + km \sum_j \delta_j + tm \sum_h \beta\gamma_{ih} + km \sum_j \beta\delta_{ij} + m \sum_h \sum_j \gamma\delta_{hj} + m \sum_h \sum_j \beta\gamma\delta_{ihj} + \sum_h \sum_j \sum_q e_{ihjq}. \quad (3)$$

From eqn. (3) the expectation of the variance over i of X_i is:

$$E(V) = k^2t^2m^2\sigma_\beta{}^2 + kt^2m^2\sigma_{\beta\gamma}{}^2 + k^2tm^2\sigma_{\beta\delta}{}^2 + ktm^2\sigma_{\beta\gamma\delta}{}^2 + ktm\sigma_0{}^2. \quad (4)$$

Now suppose that the whole experiment is repeated on a second occasion, using the same symptoms, raters and patients. Then $E(V)$ is as before, while the expectation of the covariance between the m repeats is everything in eqn. (4) except the final term, since errors, if random, should not contribute to the expected value. The expectation of the correlation between the X_i values in the experiment and in the hypothethical repeat experiment is thus given by:

$$E(\rho) = \frac{ktm\sigma_\beta{}^2 + tm\sigma_{\beta\gamma}{}^2 + km\sigma_{\beta\delta}{}^2 + m\sigma_{\beta\gamma\delta}{}^2}{ktm\sigma_\beta{}^2 + tm\sigma_{\beta\gamma}{}^2 + km\sigma_{\beta\delta}{}^2 + m\sigma_{\beta\gamma\delta}{}^2 + \sigma_0{}^2}. \quad (5)$$

If replication were possible, then all the variance components in eqn. (4) could be estimated from a set of data using the analysis of variance given in Table 4. But, in general, replication is impracticable so that $m=1$ and a direct estimate of $\sigma_0{}^2$ cannot be obtained. Consequently, the second order interaction term $\overline{DSP}$, which, if $m=1$, is an estimate of $(\sigma_0{}^2 + \sigma_{\beta\gamma\delta}{}^2)$, has to be used as an estimate of $\sigma_0{}^2$. This is equivalent to assuming that $\sigma_{\beta\gamma\delta}{}^2$ is zero—generally a reasonable assumption. Eqn. (5) now becomes:

$$E(\rho) = \frac{kt\sigma_\beta{}^2 + t\sigma_{\beta\gamma}{}^2 + k\sigma_{\beta\delta}{}^2}{kt\sigma_\beta{}^2 + t\sigma_{\beta\gamma}{}^2 + k\sigma_{\beta\delta}{}^2 + \sigma_0{}^2}, \quad (6)$$

in which $\sigma_0{}^2$ is estimated by $\overline{DSP}$. The same result is obtained from Table 4 modified to correspond to the experiment conducted without replication. First the 'replication' line is dropped. Next, all the m's in the expectations are deleted. Finally, on the assumption made in the text that $\sigma_{\beta\gamma\delta}{}^2 = 0$, this term is deleted from all the expectations. Eqn. (6) then follows at once.

An estimate of the coefficient of agreement between (i) the whole set of t raters over all k symptoms and (ii) the same set of t raters over the same k symptoms can now be obtained directly from the means squares for patients ($\bar{P}$) and the second order interaction term $\overline{DSP}$ using the equation:

$$r_1 = (\bar{P} - \overline{DSP})/\bar{P}. \quad (7)$$

Eqn. (7) is the algebraic equivalent of eqn. (6), when the estimates of the expected mean squares derived from a set of data are used.

If an estimate is required of the average coefficient of agreement of a single rater with himself over k symptoms, then we put $t=1$ in eqn. (6) and, after some algebra, obtain:

$$r_2 = \frac{\bar{P}+(t-1)\overline{DP}-t\overline{DSP}}{\bar{P}+(t-1)\overline{DP}}. \tag{8}$$

Finally, if an average coefficient of agreement of a single rater with himself on a single symptom is required, then $t=k=1$ in eqn. (6). In this case the formula in terms of estimated mean squares becomes:

$$r_3 = \frac{\bar{P}+(k-1)\overline{SP}+(t-1)\overline{DP}-(k+t-1)\overline{DSP}}{\bar{P}+(k-1)\overline{SP}+(t-1)\overline{DP}+(k-1)\,(t-1)\overline{DSP}}. \tag{9}$$

All the eqns. (6) to (9) relate to self-consistency—the agreement of a group of raters with itself, or of a single rater with himself. It is perhaps even more realistic to ask: what is the coefficient of agreement between one group of raters with another group, or one rater with another rater?

Returning to eqn. (4), let us assume that the experiment is repeated using the same symptoms and patients, but a different random sample of t raters. Then $E(V)$ is as before, while the expectation of the covariance between the repeats is the first two terms in eqn. (4), the other terms with δ's explicitly or implicitly in the subscript being excluded. Putting $m=1$, and again making the assumption that $\sigma_{\beta\gamma\delta}{}^2=0$, we obtain:

$$E(\rho) = \frac{kt\sigma_\beta{}^2+t\sigma_{\beta\gamma}{}^2}{kt\sigma_\beta{}^2+t\sigma_{\beta\gamma}{}^2+k\sigma_{\beta\delta}{}^2+\sigma_0{}^2}. \tag{10}$$

From eqn. (10) and the expectations of Table 3 we obtain at once an estimate of the coefficient of agreement between (i) the whole random sample of t raters over all k symptoms, and (ii) a further random sample of t raters over the same k symptoms. It is:

$$r_4 = (\bar{P}-\overline{DP})/\bar{P}. \tag{11}$$

An estimate of the average coefficient of agreement between one randomly selected rater and another over all k symptoms is obtained after putting $t=1$ in eqn. (10). In terms of mean squares this estimate is:

$$r_5 = (\bar{P}-\overline{DP})/\{\bar{P}+(t-1)\overline{DP}\}. \tag{12}$$

Finally, if an estimate is required of the average coefficient of agreement between one randomly selected rater and another over a single symptom, we put $t=k=1$ in eqn. (10). In terms of mean squares this estimate is:

$$r_6 = \frac{\bar{P}-\overline{DP}+(k-1)\,(\overline{SP}-\overline{DSP})}{\bar{P}+(k-1)\overline{SP}+(t-1)\overline{DP}+(k-1)\,(t-1)\overline{DSP}}. \tag{13}$$

5. THREE-WAY ANALYSIS OF VARIANCE MODEL—CASE II

Case II will be considered in the first instance for t raters, n patients, k symptoms and m replications and then n will be made unity to provide the solution required for a single patient. The basic equation is again eqn. (2) but the roles of ' patients ' and ' symptoms ' will now be reversed and, by summing over j, i and q, the total score X_h for symptom h obtained. This leads to the following expression for the expectation of the variance of the symptom totals, namely,

$$E(V)=t^2n^2m^2\sigma_\gamma{}^2+tm^2n^2\sigma_{\gamma\delta}{}^2+t^2m^2n\sigma_{\beta\gamma}{}^2+tm^2n\sigma_{\beta\gamma\delta}{}^2+tmn\sigma_0{}^2. \tag{14}$$

If the hypothetical repeat of the investigation is taken to involve a further random sample of t raters, the expectation of the covariance is the sum of the first and third terms in eqn. (14) and the expected value of the coefficient of agreement between the two samples of t raters is:

$$E(\rho)=\frac{tnm\sigma_\gamma{}^2+tm\sigma_{\beta\gamma}{}^2}{tnm\sigma_\gamma{}^2+tm\sigma_{\beta\gamma}{}^2+mn\sigma_{\gamma\delta}{}^2+m\sigma_{\beta\gamma\delta}{}^2+\sigma_0{}^2}, \tag{15}$$

which with $m=1$ and $\sigma_{\beta\gamma\delta}{}^2=0$ gives the mean square estimate:

$$r=(\bar{S}-\overline{DS})/\bar{S}. \tag{16}$$

For the particular situation in which t raters rate a *single* patient on k symptoms, $m=1$, $n=1$ and $\sigma_{\beta\gamma\delta}{}^2=0$ in eqn. (15). The expected value of the coefficient of agreement between raters now becomes:

$$E(\rho)=\frac{t(\sigma_\gamma{}^2+\sigma_{\beta\gamma}{}^2)}{t(\sigma_\gamma{}^2+\sigma_{\beta\gamma}{}^2)+\sigma_{\gamma\delta}{}^2+\sigma_0{}^2}, \tag{17}$$

which can be estimated by

$$r=\frac{\bar{S}-\overline{DS}+(n-1)\,(\overline{SP}-\overline{DSP})}{\bar{S}+(n-1)\overline{SP}}. \tag{18}$$

For the situation in which *one* rater rates a *single* patient on k symptoms $m=1$, $n=1$, $t=1$ and $\sigma_{\beta\gamma\delta}{}^2=0$ in eqn. (15). The expected value of the coefficient of agreement between two randomly selected raters is now:

$$E(\rho)=\frac{\sigma_\gamma{}^2+\sigma_{\beta\gamma}{}^2}{\sigma_\gamma{}^2+\sigma_{\beta\gamma}{}^2+\sigma_{\gamma\delta}{}^2+\sigma_0{}^2}, \tag{19}$$

for which the estimate is

$$r=\frac{\bar{S}+\overline{DS}+(n-1)\,(\overline{SP}-\overline{DSP})}{\bar{S}+(n-1)\overline{SP}+(t-1)\overline{DS}+(n-1)\,(t-1)\overline{DSP}}. \tag{20}$$

If only one patient is available for the experiment, a two-way analysis must be used in place of the previous three-way one. This restricted analysis will be of the form depicted in Table 1 with ' Patients ' replaced by ' Symptoms '.

In this case, the expected value of the coefficient of agreement between two random samples of k raters is:

$$E(\rho)=\frac{t\sigma_\gamma{'}^2}{t\sigma_\gamma{'}^2+\sigma_0{'}^2}, \tag{21}$$

which can be estimated by

$$r=\frac{\bar{S}'-\overline{DS}'}{\bar{S}'}, \tag{22}$$

where $\bar{S}'$ and $\overline{DS}'$ are mean squares in this two-way analysis.

The coefficient of agreement between two randomly selected raters is obtained by putting $t=1$ in eqn. (21). The expected value is:

$$E(\rho)=\frac{\sigma_\gamma{'}^2}{\sigma_\gamma{'}^2+\sigma_0{'}^2}, \tag{23}$$

which can be estimated by

$$r=\frac{\bar{S}'-\overline{DS}'}{\bar{S}'+(t-1)\overline{DS}'}. \tag{24}$$

This two-way analysis is less satisfactory than the previous three-way analysis. Variance components appearing in eqns. (17) and (19) are confounded in eqns. (21) and (23). Thus $\sigma_\gamma{'}^2=(\sigma_\gamma{}^2+\sigma_{\beta\gamma}{}^2)$ and $\sigma_0{'}^2=(\sigma_{\gamma\delta}{}^2+\sigma_0{}^2)$. Information is lost because the two-way analysis fails to isolate variance components which may be of interest.

6. Testing for a Bias between Raters

Returning to eqn. (2), it is also possible to find the total score for the jth rater by summing over i, h, and q. The expectation of the variance of rater totals is found to be

$$E(V)=m^2k^2n^2\sigma_\delta{}^2+m^2kn^2\sigma_{\gamma\delta}{}^2+m^2k^2n\sigma_{\beta\delta}{}^2+m^2kn\sigma_{\beta\gamma\delta}{}^2+mkn\sigma^2. \tag{25}$$

This expectation is, of course, mkn times the expected value of the mean square for $\bar{D}$ in Table 4.

(i) If the investigation is assumed to be repeated over the same t raters, the same n patients, and the same k symptoms, the expected value of the error variance of a rater total is $mkn\sigma_0{}^2$. The test for rater bias is $F=\bar{D}/\bar{E}$ where, on a similar argument to that used previously, $\bar{E}$ is taken to be the second order interaction mean square.

(ii) If the investigation is assumed to be repeated over the same t raters, a different sample of n patients, and the same k symptoms, the expected value of the error variance of a rater total becomes:

$$(m^2k^2n\sigma_{\beta\delta}{}^2+m^2kn\sigma_{\beta\gamma\delta}{}^2+mkn\sigma_0{}^2).$$

The test for rater bias is then $F=\bar{D}/\overline{DP}$.

(iii) If the investigation is assumed to be repeated over the same t raters the same n patients, and a different sample of k symptoms, the expected value of the error variance of a rater total now becomes:

$$(m^2kn^2\sigma_{\gamma}^2+m^2kn\sigma_{\beta\gamma\delta}^2+mkn\sigma_0^2),$$

and the test for rater bias is $F=\bar{D}/\overline{DS}$.

(iv) If the investigation is assumed to be repeated over the same t raters, a different sample of n patients, and a different sample of k symptoms, the expected value of the error variance of a rater total is now

$$(m^2k^2n\sigma_{\beta\delta}^2+m^2kn^2\sigma_{\gamma\delta}^2+m^2kn\sigma_{\beta\gamma\delta}^2+mkn\sigma_0^2).$$

The test for rater bias is $F=\bar{D}/(\overline{DS}+\overline{DP}-\overline{DSP})$. The degrees of freedom ar troublesome, but if n and k are fairly large, a decision about the significance of would be sufficiently clear-cut.

7. Dichotomously Scored Data

Consideration is now given to situations in which rating scales are not employed but the raters simply indicate whether a symptom is present or absent scoring these 1 or 0 respectively. One method of dealing with Case I*a* is t apply the well-known Kuder–Richardson formula KR 20 (DuBois, 1965).

Suppose t raters $(i=1, 2, \ldots, t)$ each rate all the patients for the sam symptom. The proportions rated 1 and 0 by rater i are p_i and q_i respectively, and the variance of patient totals over all t raters is s^2. Then an estimate of the correlation between these patient totals and the corresponding totals over a parallel set of t other raters is

$$r_{tt}=\frac{t}{t-1}(s^2-\sum_i p_iq_i)/s^2. \tag{26}$$

KR 20 is readily derived from a two-way analysis of variance in which the sources of variation are 'patients' and 'raters' and the variate consists of 1's and 0's denoting the presence or absence of a particular symptom.

Eqn. (26) is also Cronbach's 'coefficient alpha' (Cronbach, 1951) in the special case of dichotomous rating. In the present context, writing $\bar{W}$ for the mean square and s_i^2 for the mean variance 'within raters', it is found that

$$\alpha=\frac{t}{t-1}(\bar{P}-\bar{W})/\bar{P}=\frac{t}{t-1}(s^2-\sum_i s_i^2)/s^2,$$

from which eqn. (26) follows at once.

An estimate of the average agreement between one rater and another is:

$$r_{11}=(s^2-\sum_i p_iq_i)/\{(t-1)\sum_i p_iq_i\}. \tag{27}$$

The Kuder–Richardson formula cannot be employed for dichotomously scored data falling in Cases I*b* or II. In these cases the analysis of variance methods outlined in Sections 3 and 4 above can be used, provided the assumptions underlying the analysis are reasonably fulfilled. In particular, analysis of variance models would be apposite when the incidence of the symptoms, as recorded by each rater, is within the range 20–80 per cent for then the variances of the symptoms would meet the criterion of homogeneity fairly adequately.

More recently, Case I*a* (and, by repeated application, Case I*b*) has been discussed by Armitage *et al.* (1966) and by Fleiss (1965). Armitage *et al.* adopt a straightforward mathematical approach to the problem in terms of proportions and propose three indices of observer agreement or disagreement, any one of which might be used. They are: (*a*) The mean Majority Agreement Index, measuring the frequency with which observers agree with the majority opinion; (*b*) The mean Pair Disagreement Index, measuring the frequency with which pairs of observers disagree; (*c*) The Standard Deviation Agreement Index, measuring the variation between subjects in the number of observers. Use of any of these indices will tend to yield similar conclusions; (*b*) and (*c*) are particularly closely connected. The authors add the cautionary note that ' the numerical value to be expected for any index depends strongly on the proportion of positive findings in the whole study '. Fleiss (1965) feels that an attack on the problem in terms of simple proportions may not always be adequate (but, unlike Armitage *et al.*, he has tests of significance and confidence limits in mind). In particular, he wishes to avoid any assumption about normality of errors when dichotomous judgements are made. Consequently, he sets up a non-parametric model for dealing with errors underlying such judgements. Using the model, he proceeds to give conditions under which ' Cochran's Q statistic is valid for testing the hypotheses of no systematic differences amongst the judgements of the different observers '. He also provides a measure of intraclass correlation between all pairs of judgements on the same subject which may be used as a measure of concomitant agreement between observers in cases where systematic differences between them are small.

REFERENCES

ARMITAGE, P., BLENDIS, L. M. & SMYLLIE, H. C. (1966). The measurement of observer disagreement in the recording of signs. *J. R. statist. Soc.* (*A*) **129**, 98–109.
BARTLETT, M. S. (1947). The use of transformations. *Biometrics* **3**, 39–51.
BOX, G. E. P. & COX, D. R. (1964). An analysis of transformations. *J. R. statist. Soc.* (*B*) **26**, 211–243.
CRONBACH, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297–334.
DUBOIS, P. H. (1965). *An Introduction to Psychological Statistics*. New York: Harper & Row.
FLEISS, J. L. (1965). Estimating the accuracy of dichotomous judgements. *Psychometrika* **30**, 469–479.
FLEISS, J. L. (1966). Assessing the accuracy of multivariate observations. *J. Am. statist. Ass.* **61**, 403–412.

FLETCHER, C. M. (1960). Criteria for diagnosis and assessment in clinical trials. In A. Bradford Hill (ed.), *Controlled Clinical Trials*. Oxford: Blackwell.

HAMILTON, M. (1959). The assessment of anxiety states by rating. *Br. J. med. Psychol.* **32,** 50–55.

HOYT, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika* **6,** 153–160.

JACKSON, R. W. B. (1940). *Application of the Analysis of Variance and Covariance to Educational Problems*. Toronto: University of Toronto.

MCNEMAR, Q. (1962). *Psychological Statistics*. New York: Wiley.

WINER, B. J. (1962). *Statistical Principles in Experimental Design*. New York: McGraw-Hill.

Vol. 21 Part 1 117–130 The British Journal of Mathematical and Statistical Psychology May 1968

SIMULATING PERCEIVED SIMILARITIES BETWEEN TASTE MIXTURES HAVING MUTUALLY INTERACTING COMPONENTS

By R. A. M. Gregson
University of Canterbury, New Zealand

Empirical findings about the perceived interaction of the components of qualitatively complex taste stimuli are reviewed and summarized. The revision of multidimensional scaling equations to incorporate a simulation of a mutually inhibitory or facilitatory interaction between components is discussed. Data from a previously published study on the perception of similarity among taste mixtures are re-analysed for illustrative purposes and a description of them is shown to be less adequate when provision is made for one form of interaction than when this is omitted. A reclassification of stimulus–response paradigms for studying taste mixture perception is outlined; and it is suggested that the interaction of taste components may be an artifact of the response measures used.

1. Introduction

When a mixture of qualitatively different tastes is experienced by a human subject, it is sometimes reported that the components of the mixture interact. For example, if sucrose of a given concentration is present, the perceived intensity of sweetness of the mixture may be more, or less, than would be the case if the sucrose were in a simple solution with no admixture of other components which, taken singly, elicit qualitative sensations other than sweetness. Interaction, in this context, is thus an approximate psychophysical analogue of interaction in analysis of variance models, in the sense that the set of qualitatively distinct intensities resulting from a complex stimulus cannot be predicted from a linear sum of the intensities which result when the components of the complex stimulus are experienced one at a time.

The problem has been of recurrent interest to workers in physiology, psychology and food technology, because the phenomena of sensory interaction between taste components are the stumbling-block for any theory of taste, and particularly for any multidimensional scaling models of taste perception which aim to map more than a limited part of between-judgement variance. The measurement theory problems which arise when an attempt is made to represent multidimensional systems involving interactions are not, of course, unique to taste. However, in the case of taste the problems are more obvious and perhaps more acute.

Empirically, the problem of taste mixture perception has been studied piecemeal for some time (Bujas, 1934; Fabian & Blum, 1943; Beebe-Center *et al.*, 1959; Pangborn, 1960, 1962; Kamen *et al.*, 1961; Gregson & McCowen, 1963; Békésy, 1964; Gregson, 1966 *a*; Indow, 1967). Many different chemical stimuli and a variety of response measures and psychophysical paradigms have been employed. The results are not consistent from study to study, mainly due to methodological differences in stimulus presentation and in the way subjects may be restricted to one particular form of report concerning the qualitative and intensity changes of one sensory component in a mixture. Both absolute and relative intensity ratings have been used.

Consistent and reliable differences between subjects in the direction of sweet–acid interactions (acid making sucrose more sweet or less sweet) have been reported in more than one of the cited studies; results on other stimulus pairings are less clear-cut. Similar directional differences apparently may depend on a change of method from eliciting absolute ratings of the component intensity of a taste mixture to eliciting relative intensity ratings of that component against the same qualitatively identical concentration in an unadulterated mixture. The diversity of results suggests that the specification of response constraints, or the related and equally difficult specification of what the subject is to judge in the task, is likely to be as critical as the description of the stimulus. The earlier studies have been extensively reviewed by Amerine, Pangborn & Roessler (1965) but little, if any, attention has been given to the psychophysics of the rating task as compared with the amount of speculation there has been about the chemistry of the receptor cell environment when complex stimuli are employed.

However, the general finding, possibly subject to minor qualification for some taste pairings, is that the presence of one component tends to diminish the perceived intensity of some other component which yields by itself a qualitatively distinct sensation. For example, adding quinine sulphate to sucrose both introduces a bitter taste to a mixture and makes the mixture less sweet than would be the case if the quinine sulphate were not present. (There is a phenomenological problem here in that some subjects cannot judge sweetness and bitterness separately but coalesce them into one bipolar sensation. This fact has been employed by Békésy (1964) in a duplexity theory of taste.) The extent of the inhibitory effect increases with the amount of the inhibiting component in relative terms, whilst in absolute terms the extent of the inhibition can be approximated by a positive monotonic function of the ratio of the concentration of the inhibited to the sum of the inhibiting components, given an appropriate response metric.

This relatively simple type of interaction, which if mutually inhibitory is conveniently referred to as ‘ buffering ’, is not the only type which has to be considered. In all psychophysical representations of taste perception a partial isomorphism between the sets {sweet substances, most acids, some salts, bitter substances} (usually exemplified by {sucrose or glucose, hydrochloric or citric acid, sodium chloride, caffein or quinine sulphate}) and {sweetness, sourness, saltiness and bitterness} is assumed to operate at least over any range of concentrations and intensities which does not include the near-threshold region on any single dimension. In the near-threshold region difficulties arise because the qualitative sensation changes as a function of concentration.

2. Multidimensional Representation of Interactions

Interaction could arise as readily from adaptation or frame-of-reference effects in the judgement process (Gregson & Paris, 1967) as from any receptor cell events. But, whatever its causal locus, provision has to be made for its possible occurrence in any psychophysical model of taste perception which is claimed to be valid. Indeed, by using multidimensional scaling, it might be thought easier to arrive at a general assessment of the form or magnitude of inter-dimensional inhibitory effects than would be the case from conducting a series of experiments in which each possible pair and trio of interactions was studied separately. Given sweet, acid, salty and bitter as sensory dimensions there are ten such subsystems, $\binom{4}{2}+\binom{4}{3}$, neglecting differences between stimulus substances eliciting the same primary sensation.

Whilst the near-threshold region can usually be avoided in simple unidimensional psychophysics, it is almost certain to be involved in any stimulus series employed in multidimensional scalings on at least some dimensions; without such weak intensities on some dimensions in some of the mixtures employed, an even spacing of mixtures in a stimulus hyperspace is unlikely to be achieved (Gregson, 1965). The locus of the effects of near-threshold qualitative changes in the multidimensional scaling experiment could be predicted if more extensive data on this breakdown were available than is the case at present, because such experiments normally employ a complete paired comparison matrix and not all the cells in such a data matrix are equally susceptible to departures from the predictions of a scaling model which fails to make provision for the breakdown of sensory isomorphism.

Multidimensional scalings of taste mixture similarity perception (Gregson, 1965, 1966 *b*, *c*; Gregson & Russell, 1965 *b*; Russell & Gregson, 1966) have so far made no explicit provision for interaction. This paper describes a model with such provision, and discusses the results and problems arising from its use.

Let a taste stimulus x have t dimensions $\{i=1, 2, \ldots, t\}$, and let the physical stimulus magnitude on the ith dimension be x_i for the mixture $x \in M$, where M is a set of m taste mixtures between which comparisons are to be made. The total sensory input for x is then the set $\{x_1, x_2, \ldots, x_i, \ldots, x_t\}$.

If the input is of the form $\{0, 0, \ldots, x_i, 0, \ldots, 0\}$, that is to say a mixture with one non-zero component, then the response—a perceived intensity—will be $\psi_i(x_i)$, where ψ_i is the operator effecting the unidimensional psychophysical transformation to the sensory dimension which corresponds to stimulus dimension i. Various ψ may be merely scalar multiples of each other; e.g. if more than one x_j is a salt, the relevant ψ_j are all transformations to 'saltiness' which are known to be empirically similar (McBurney & Lucas, 1966). In general, following Stevens (1961), and using i_0 as a zero level symbol for dimension i,

$$\psi_i(x_i) = \alpha_i(x_i - i_0)^{s_i}, \tag{1}$$

where α and s are psychophysical constants. What evidence exists suggests that s_i is about unity for each of the four primary taste dimensions, and as a simplifying assumption s will be written for all s_i hereafter, and, where numerical best *a priori* guesses are needed, s will be set equal to unity. This is compatible with some earlier versions of taste scaling (Beebe-Center & Waddell, 1948; Dove, 1953). Hardly any reliable data on differences between s_i as compared with inter-individual differences on any one s_i have been published (Gregson & Russell, 1965 *a*). In the mixture x^*, $x^* \in M$, of the form $\{x^*_1, x^*_2, 0, 0, \ldots, x^*_i, 0, \ldots, 0\}$, that is to say with, arbitrarily, three non-zero components one of which, x^*_i, has the same intensity as x_i, then

$$\psi_i(x^*_i) = \alpha_i(x^*_i - f(x^*_1, x^*_2, x^*_i) - i_0)^s, \tag{2}$$

where f is positive monotonic on x^* component intensities. As $s > 0$, $x^*_i = x_i$, and the buffering effect f is always less than or equal to the component it is buffering (one cannot buffer sweetness below zero intensity into 'negative sweetness', whatever that might be), so that

$$x^*_i - i_0 \geqslant f(x^*_1, x^*_2, x^*_i) \geqslant 0,$$

then in general

$$\psi_i(x^*_i) < \psi_i(x_i).$$

(If $s=2$ this inequality may not be correct, but it has never been found to be higher than 1·3 in empirical studies.)

Following earlier studies and results (Gregson, 1966 *b*, *c*, *d*; and unpublished data by Russell), the psychological distance between two mixtures x, y may be represented to a first approximation by

$$\psi(x, y) \Leftrightarrow D^*(x, y\colon \{t\}, r) = \left[\sum_{i=1}^{t} (\alpha_i \mid (1-\theta_i)(x_i - \epsilon_i) - (y_i - \epsilon_i) \mid^s)^r\right]^{1/r}, \tag{3}$$

where $|\theta_i| < 1$, $\epsilon_i \ngtr \min x_i, y_i$ for all t dimensions; r corresponds to the Minkowski r parameter in a generalized metric of distance, and θ is a parameter representing time-order error or trace decay. Eqn. (3) is to be read as giving the strict order of comparison: stimulus x always precedes y. The parameters ϵ_i obviously correspond formally to the i_0 of eqn. (1). The change in notation is intended to draw attention to the fact that i_0 is always defined for a dimension occurring in isolation, and ϵ_i is estimated on the assumption that it will not be very different from i_0, but need not be the same because of the presence of collateral stimulation. If taste dimensions can be confused, phenomenologically or otherwise (it is known that bitter and sour are readily confused semantically, and particularly at low intensities), then an effective threshold shift specific in its action and magnitude to a single dimension may arise and need representation in this way. This form of eqn. (3), with appropriate modification for cross-modal comparisons, has been used (Gregson, 1966 *d*) because it appears to fit the data slightly better than earlier versions (Gregson, 1966 *b*).

Eqn. (3) can be modified to take account of mutual inhibition in a manner consistent with its supposed form if the simplest published data is accepted at its face value. A term of the following kind is introduced:

$$\beta[\text{buffering components/buffered component}], \tag{4}$$

where β is a constant representing the extent of buffering effects. The buffering components are to be represented by the sum of their effective intensities. Expression (4) will be regarded as subtracting from x_i and y_i, and becomes an additive but regular component summating with the effects of the ϵ_i. The x_i are being thus reduced by three conceptually distinguishable additive effects: (i) the unidimensional detection threshold, (ii) a buffering due to effects which can be expressed in the general form (4), and (iii) residual buffering effects specific to individual dimensions, over and above those accounted for by expression (4). The two effects (i) and (iii) are represented by the ϵ_i.

Eqn. (4) can be rewritten as

$$Bx_i = \beta\alpha_i^{-1}(x_i-\epsilon_i)^{-1}\sum_{j=1}^{t-1}\alpha_j(x_j-\epsilon_j), \tag{5}$$

where $j \neq i$. By_i is similarly defined. In eqn. (5) the notation indicates summation over $t-1$ dimensions excluding the ith, and not summation up to the $(t-1)$th dimension.

Combining the Bx_i and By_i terms in eqn. (3) yields the fullest form of interaction equation to be considered here:

$$\psi(x, y) \Leftrightarrow D^{**}(x, y; \{t\}, r) = \left[\sum_{i=1}^{t}(\alpha_i \,|\, (1-\theta_i)\,(x_i-\epsilon_i-Bx_i)-(y_i-\epsilon_i-By_i)\,|^{s})^{r}\right]^{1/r}, \tag{6}$$

which is in $(3t+3)$ unknowns and predicts all the m^2 comparisons over an $M \times M$ experiment. Hence, for an adequately determined solution with enough degrees of freedom, $m^2 > (3t+3)$, which is readily satisfied. To find a solution set of parameters for α, β, ϵ, θ, s, and r for a given experiment, it was decided first to solve eqn. (3) iteratively, which is, in fact, eqn. (6) with $\beta = 0$, and then gradually move β away from zero in either direction. This strategy exploits the fact that interaction effects reported are definitely second-order. If they were not, there would be a strong risk of getting a different solution this way from that obtained when proceeding directly to solve eqn. (6) by iterative methods. There is a problem here, in any case, for the form of eqn. (6) does not exclude the possibility of multiple solutions to the data matrix matching by virtue of some sort of 'trading relation' between the effects of parameters β, and θ and ϵ together, and it seems that only by using psychophysical knowledge about the relative expected magnitudes of the effects mapped into eqn. (6) can the relative merits of two solutions be decided.

Two comments are pertinent: in some experiments the x_i have an approximately constant sum for all $x \in M$, a precaution deliberately introduced to avoid certain adaptation effects in the scaling task. In this case the low x_i might be heavily inhibited and the high x_i lightly inhibited. This in turn implies that

$\psi_i(x_i)$ rises more steeply, as a consequence of inhibition, than it would otherwise. This in turn demands representation in eqn. (3) by a higher Minkowski r. So, in a mixture set with fixed constant effective overall intensity for each mixture in the set it would be expected that if a real positive β term should be present, its absence would be compensated for, in part, by a raised r. Conversely, if we introduce a β term and raise it too high, r should drop to compensate in order to maintain a good fit to empirical data. This excludes testing eqn. (6) against data from experiments where the constraint on the total intensity of mixtures has been effectively applied.

Secondly, the assumption that β is a constant over all i, j dimension pairs is a simplification. If a different value is assumed for each i, j pair, in each direction of inhibition (i upon j, or j upon i) then the problem becomes intractable because (t^2-t) unknowns are added and iterative solutions become interminable and untrustworthy. More simply, β might be replaced by a product term $\beta_i \times \beta_j$, where the first constant is associated with the inhibited component and the second with the inhibiting component. This possibility is not developed here because the available empirical evidence does not justify the effort in the context of taste, but it is advanced because it might be worth considering in intersensory studies, as opposed to interdimensional studies of interaction within a modality.

It is here proposed to solve eqn. (6) by putting

$$H=[\psi(x, y)-D^{**}(x, y)]^2 \tag{7}$$

and finding a stationary value for H, by a Monte Carlo method. Such a value is reached when

$$dH=\sum_{i=1}^{t}\frac{\partial H}{\partial \alpha_i}\cdot d\alpha_i+\sum_{i=1}^{t}\frac{\partial H}{\partial \theta_i}\cdot d\theta_i+\sum_{i=1}^{t}\frac{\partial H}{\partial \epsilon_i}\cdot d\epsilon_i + \frac{\partial H}{\partial \beta}\cdot d\beta+\frac{\partial H}{\partial r}\cdot dr+\frac{\partial H}{\partial s}\cdot ds=0. \tag{8}$$

To simplify matters, s is dropped. Unfortunately, there are no immediately appealing constraints that might be imposed upon eqn. (8) to facilitate analytical solution by, say, Lagrangian undetermined multipliers (Freeman, 1963) but there are the following limits on acceptable solutions.

The term in α_i can be over $t-1$ dimensions, for any one α_i can be put equal to unity and the rest are relative values. Additionally, a notional maximum (or minimum) on any α_i can be imposed, say $|\alpha^*|$, and similarly the maxima $|\theta^*|$ and $|\epsilon^*|$ can be assumed from unidimensional psychophysics. It has been found that α_1 (for sweetness) is generally larger than the remaining α_i. These α_i are actually weights proportional to the inverse variance of the data on the dimensions which they represent (assuming the units of measurement are comparable between dimensions, which depends on the actual solution concentrations chosen). The larger an α_i, the more precise the estimates on i. It is therefore convenient to use as trial values for the α_i the relative mean inverses of the Weber fractions on the four taste dimensions involved over the m^2 paired comparisons. These values will be predicted to be nearly equal, and have been found to be so (Gregson, 1966 *d*). The product of the α_i is a function of the volume of space in which the solution to eqn. (7) is contained; the limits of this volume are given from

$$1 \leqslant \prod_{i=1}^{t} \alpha_i \leqslant |\alpha^*|^3,$$

as shown by Sebestyen (1962).

Stationary values of H generally occur on a mesa (the surface of a flat-topped mountain) as Minsky (1961) has previously observed. This hypersurface of constant H-altitude has a perimeter given by the tolerance, w, on each of the α, θ, ϵ, β, r. The vector $\{w\alpha_i, w\theta_i, w\epsilon_i, \ldots\}$ $(i=1, \ldots, t)$ sweeps out the mesa surface about the final adopted solution. The partials of eqn. (8) are slopes of ascent to the mesa, and zero on the mesa itself. For example, the surface runs on dimension α_i from $\alpha_i - w'\alpha_i$ to $\alpha_i + w''\alpha_i$, where w' and w'' (lower and upper tolerances) are not necessarily equal; $w\alpha_i$ summarizes this state of affairs.

The location of the mesa surface for min H in the parameter hyperspace may be modified by the introduction of β into eqn. (6) in three ways, neglecting the additional term w in the vector: (i) no difference, and $\beta = 0$; (ii) the altitude and bounds of the mesa are moved, and $\beta \neq 0$; (iii) the altitude stays the same, the bounds are moved, and $\beta \neq 0$. Case (i) corresponds to the solution reported here. Case (ii) would correspond to an improvement or a worsening consequent upon mapping the reported forms of buffering into the model. Ideally, the mesa would shrink to a single peak, so improvement may be conceived of as either an elevation or an area reduction for the mesa. If elevation consequent on $\beta \neq 0$ also entailed an area increase, this would suggest an undesirable inter-dependence of parameter estimation. Case (iii) is really two cases; either the altitude remains the same but the solution set of parameters and the bounds are moved with the area the same or greater—which is no improvement and suggests the absence of a 'best' solution—or the altitude remains the same but the mesa area is reduced with or without a change in the solution set of parameters excluding β. Such a reduction is an improvement as the information in the solution in Fisher's (1934) sense is increased, and solutions could then be ordered on information as well as on goodness of agreement for H. Overall, the requirement must be imposed that the location of the mesa surface where iteration or exploration finishes is independent of the path by which the mesa is reached (cf. Wilde, 1964).

Empirically, min H is approached quickly by altering α_i values cyclically in steps, then altering θ_i or ϵ_i, and this procedure has been used. It obviously has risk as the price of speed. When any $\partial H/\partial\alpha_i$ is not near zero it swamps the effects of the $\partial H/\partial\theta_i$ because α_i multiplies θ_i. When the mesa is neared, the $\partial H/\partial\alpha_i$ are near zero; for the altitude to be increased, $\partial H/\partial\beta > 0$, which is a condition of independence between β and α_i which could permit a better solution.

In the methods of Shepard (1962 *a*, *b*) and Kruskal (1964), x_i and y_i are the values modified by iterations (apart from r), and are not, as here, fixed *a priori* whilst a smaller number of parameters is adjusted. Instead of a mutually interdependent configuration of coordinate vectors, we have a set of parameters which may be independently varied within limits, each modifying the effective coordinates of one or more points.

In all the multidimensional scaling methods advanced, the possibility of false solutions (ending up on the wrong mesa) have been commented on by authors, and found in practice. The final values for x_i are not mutually independent in Shepard's or Kruskal's methods of non-metric scaling as the degrees of freedom implicit in their monotonicity (of ψ on D) requirements are smaller than the number of points for minimum dimensionality. Here, the degrees of freedom are greater than the number of parameters, and dimensionality, t, is fixed *a priori* on the basis of previous results. The problem here is whether or not there exists a mesa with $\beta \neq 0$ and bounds not coincident with the $\{w\alpha_i, w\theta_i, w\epsilon_i\}$ implied here, *and* higher altitude: to search for this in hyperspace is equivalent in effort to searching deliberately for as many false solutions as may exist. With very high speed computation and a series of starts from within $\{|\alpha^*|, |\theta^*|, |\epsilon^*|, |\beta^*|\}$, this is notionally testable, but the plausible bounds on $|\beta^*|$ are such that, unless there are a number of steep-sided local mesas, the $\partial H/\partial\alpha_i$ would generally swamp the effects of $\partial H/\partial\beta$ and lead back to the mesa partially tabulated below.

In fact, the precise location of a best solution may be viewed as irrelevant to the question of whether or not buffering should be mapped into the model, given the limited volume of parameter space within which a solution may be sought. One may set up, as competing Bayesian hypotheses (Hunt, 1965), $\beta = 0$ and $\beta = 0{\cdot}5$, and compare the goodness of fit of a spaced sample of parameter combinations inside the plausible limits. The dominance of one set of solutions over the other is taken as supporting evidence for, say, $\beta = 0$ as against $\beta = 0{\cdot}5$. Dominance here means that the correlation r for $\beta = 0$ is greater than that for $\beta = 0{\cdot}5$ in all parameter sets examined, or in nearly all, including all in the subset of cases having high r for both $\beta = 0$ and $\beta = 0{\cdot}5$. The *a posteriori* probability of the dominated hypothesis rapidly approaches zero as the number of sets examined increases. It must be emphasized that this procedure, however, only eliminates the hypothesis of a general regular buffering with $\beta = 0{\cdot}5$; it does not eliminate the alternative having residual positive and negative ϵ_i and a nearly zero mean buffering effect, over all $\binom{4}{2}$ possible interactions between dimensions.

The D^{**} equation, modified to apply to cross-modal matching, has been applied to results taken from a study involving four-component taste mixtures matched to four-component histograms (Gregson, 1966 *d*). Five mixtures of varying proportions of sucrose, citric acid, sodium chloride and quinine sulphate were cross-modally matched to eight four-column histograms, where the columns were colour-coded to represent the tastes: sweet, acidic, salty and bitter. The resultant matrix of $5 \times 8 = 40$ cross-modal inter-stimulus distances was generated by 16 subjects in each of four conditions, described in the original paper. The similarities between taste mixtures and the mixtures represented by the histograms were rated by the subjects on a scale from 0 (' identical ') to 8 (' impossible for the two mixtures to be the same '). Considering one condition only, that designated High Sucrose, Unpaced Matching in the original study, the following results relevant to the present paper were obtained.

The overall mean similarity on the scale, over 40 comparisons, was 4·49; the variance of the 40 cell means in the $M \times M'$ data matrix was 0·263; and the estimated reliability of the 40 cell means was 0·563.

The variance of the cell means is lower than in some unpublished experiments conducted more recently, and greater than in one other condition in the original study from which the data are taken. It varies with response constraints such as pacing; its absolute value is not informative as subjects do not effectively use all the response information available to them.

The best fit of eqn. (3), modified for the cross-modal case by subtracting a threshold ϵ_i from each x_i value and solving for all ϵ_i, and putting $s = 1$, gave a product-moment correlation between the observed and theoretical similarities matrices of 0·807, with 25 d.f. Parameter values selected, all understood to have a tolerance of about $\pm 0{\cdot}1$ for α and ϵ, $\pm 0{\cdot}02$ for θ, with Minkowski $r = 2{\cdot}0$ were:

$$\begin{array}{lll}
\alpha_1 = 1{\cdot}8 & \epsilon_1 = -2{\cdot}4 & \theta_1 = -0{\cdot}16 \\
\alpha_2 = 0{\cdot}8 & \epsilon_2 = 0{\cdot}4 & \theta_2 = -0{\cdot}08 \\
\alpha_3 = 1{\cdot}0 & \epsilon_3 = 0{\cdot}2 & \theta_3 = 0{\cdot}03 \\
\alpha_4 = 0{\cdot}9 & \epsilon_4 = 0{\cdot}4 & \theta_4 = 0{\cdot}06
\end{array}$$

The negative θ_1 and θ_2 values are unexceptional; time order effects in taste are variously positive or negative.

Eqn. (6) was modified for the cross-modal case by putting the values of By_i and the ϵ_i subtracted from y_i both equal to zero. In the cross-modal case the y_i values always refer to the stimulus presented visually; and it is assumed that there is no interaction between the histogram columns and no threshold effect which would modify their perceived relative lengths.

The amount of variance accounted for by this scaling is not very high compared with results in unidimensional psychophysics. The extent to which the variance of data can be accounted for by scaling models seems to be reduced as the number of dimensions goes up, and is probably always low for the chemical senses. Similar findings in this respect have been obtained for more conventional multidimensional procedures by Ekman & Engen (1962).

The solutions to the equations for the 40 cross-modal comparisons were calculated on an IBM 1620 computer. The parameter values which were explored with the α, ϵ and θ values given above are set out in Table 1. It is obvious that progressive modifications of β away from zero in either direction produced a progressive and marked drop in the correlation between the observed and the synthetic data matrices.

TABLE 1. CORRELATIONS BETWEEN OBSERVED AND SIMULATED DATA MATRICES AS A FUNCTION OF PARAMETER COMBINATIONS EXPLORED

Minkowski r	β	s	*Product moment* r	*Minkowski* r	β	s	*Product moment* r
2·0	−0·2	1·0	−0·273	2·0	0·0	1·2	0·802
	−0·1		−0·050		0·05		0·726
	−0·05		0·596		0·1		0·653
	−0·03		0·780	2·0	0·0	0·8	0·791
	−0·02		0·801		0·05		0·706
	−0·01		0·808		0·1		0·608
	0·0		0·807	2·8	0·0	0·8	0·806
	0·01		0·793		0·05		0·725
	0·02		0·793		0·1		0·650
	0·03		0·757	2·8	−0·1	1·0	−0·091
	0·05		0·723		−0·05		0·416
	0·1		0·643		−0·01		0·806
	0·15		0·570		0·01		0·778
	0·2		0·503		0·05		0·727
	0·25		0·450		0·1		0·657
				1·8	−0·2	1·2	−0·287
					−0·05		0·583
					0·0		0·806

Repeated values are not shown in this table

Using the correlation as a measure of best fit is taking a least-squares measure; the residual squared deviation about the linear relation of ψ on D^{**} is minimized. Thus a stronger requirement is imposed on the D^{**} values than a simple monotonicity with the ψ values. The decision to take this measure reflects the strength of the author's belief that D^{**} is an appropriate psychophysical function for taste.

For the case arising when all $\alpha_i = 1{\cdot}0$, and all $\theta = \epsilon = \beta = 0$, $s = 1{\cdot}0$, and the Minkowski $r = 2{\cdot}0$ (which is a simple Euclidean distance function as used by Shepard (1962 *a*, *b*)), the correlation $r = 0{\cdot}734$. The same 'basic' case, with $\beta = 0{\cdot}5$, yielded $r = 0{\cdot}738$. A range of 20 parameter combinations with unequal α_i, and some or all θ and ϵ not zero, were checked, in each case with $\beta = 0$ and $\beta = 0{\cdot}5$ as best *a priori* alternatives. This was to determine whether some alternative combinations inferior with $\beta = 0$ were better with $\beta = 0{\cdot}5$. In all these cases the solutions for $\beta = 0{\cdot}5$ were inferior to those for $\beta = 0$, the correlation r values for which lay between 0·73 and 0·79, and none of the combinations equalled the chosen set in goodness of fit. This evidence would obviously be strengthened if it were repeated over other independent experiments, but it does diminish *a posteriori* confidence in the hypothesis of buffering as it is modelled here.

3. DISCUSSION

Whilst taste component interaction is an established empirical phenomenon, no support has been found, when scaling overall similarities between mixtures (as opposed to relative intensities of single qualitative dimensions in mixtures), for a scaling model which explicitly makes provision for a specified kind of mutually inhibitory or facilitatory interaction between components. On the contrary, the best fit recorded employs parameter values which presuppose no interaction. Either interaction is unlikely, or the model does not appropriately represent interaction.

This finding raises a number of crucial problems about the application and interpretation of multidimensional psychophysics for a sensory channel which has the properties characteristic to taste. These problems are perhaps best exemplified by considering some alternative stimulus–response relations which could mediate taste mixture perception. (In each case the relation of sweetness to sucrose is taken as a paradigm.) Table 2 lists the cases which it is convenient to distinguish.

TABLE 2

Effective stimulus	*Response recorded*
(i) Absolute sucrose concentration.	Intensity of perceived sweetness.
(ii) Sucrose buffered at the receptor level by some other substance.	Intensity of perceived sweetness.
(iii) Sucrose concentration relative to the total stimulus concentration, assuming cross-stimulus intensity comparisons mediated at the receptor or peripheral level to be psychologically meaningful.	Intensity of perceived sweetness.
(iv) Sucrose concentration, *and* the concentration of each other component present, regarded as superimposed but non-interacting excitation patterns at the peripheral level.	Sweetness intensity judged relative to the total sensation level, either (*a*) as mediating processes for a final ratio judgement or (*b*) as a final absolute perceived intensity.
(v) The trace of the sensory response to the sucrose component on a first presentation, and the response to the sucrose component on a second presentation, both centrally stored.	Perceived relative intensity based on a sequential comparison of the two sensation levels centrally stored.

Cases (i) to (iv) are appropriately studied by an absolute psychophysical method. Case (v), which will also involve one of the procedures (i) to (iv), is appropriate for treatment by psychophysical methods involving relative judgement. Case (v) is also appropriate for description by the multidimensional similarity methods associated with eqns. (3) or (6), if it is assumed that the process whereby dimensions are pooled into the overall similarity response is independent of the interaction process. In eqn. (6) this assumption is made, for pooling is described by the parameter r, and interaction by β. As no constraints are imposed on the values of these two parameters, no presupposition is made that either pooling or interaction do in fact occur. For each case above, the existence or absence of interaction may be predicted to show itself as follows, where the numbers again refer to the cases described above.

(i) No difference is expected between perceived intensity responses to a given stimulus concentration, whether the stimulus is presented alone or in a mixture.

(ii) Ignoring complications near the lower threshold region, sucrose of a given concentration should appear less sweet when in a mixture, but the extent to which its sweetness is depressed is not directly predictable from the relative intensity judgements of the sucrose and other components experienced singly.

(iii) Sucrose will appear less sweet in mixtures. If x is the sucrose concentration and y the other component concentration, $\psi(x/(x+y))$ is the only possible response to x in this paradigm when it is in a mixture.

(iv) Sucrose will appear less sweet in mixtures; the extent to which it will be less sweet is predictable from the following notional experiment. If $\psi(1)$ is the intensity of a standard solution of 0·01 g./ml. sucrose, called one *gust* by Beebe-Center & Waddell (1948), then define

$$g(x)=\frac{\psi(x)}{\psi(1)}.$$

Then if $g(z)$ refers to the interacting and qualitatively not-sweet taste component experienced singly, and $g(x')=g(z)$, x' is the sucrose concentration which matches, across dimensions, the intensity of the interacting taste. Denoting the intensity of the conjoint presentation of x and z by $G(x, z)$, and the intensity of x when perceived in the context x, z by $\gamma_{x,\,z}(x)$, it is predicted that

$$\frac{g(x)}{G(x, z)}=\frac{g(x)}{g(x+x')}=\gamma_{x,\,z}(x).$$

This formalization can readily be extended to t dimensions, and the γ values substituted for x values in an equation of the form of eqn. (3).

(v) What happens in this case depends on which of the four previous mechanisms is involved prior to the comparative judgement. An 'equality' response is virtually an artifact of experimental conditions, because time–order effects make physically equal stimuli appear different when experienced sequentially.

If the comparative judgement made is an estimated similarity, and not a relative magnitude or an absolute sensory intensity, then there are no experimental precedents involving unidimensional responses to multidimensional complex stimuli which could serve as a clear basis for prediction. If the response transformations mediating similarity judgements (as opposed to the other two forms which are akin to ratio judgements) are not isomorphic with the latter transformations, then it is logically possible that interaction would not show itself within the limits of data reliability and the relative magnitude of similarity judgements. In the matrix scaling procedures previously used by Russell &

Gregson (1966), based on the earlier work of Torgerson, Shepard, and Kruskal, the raw data have been similarities which are only invariant up to a linear transformation.

4. CONCLUSIONS

The conflicting evidence which has been discussed does not make interaction implausible, but it does suggest that interaction as usually reported (psychophysically and not physiologically at the receptor mechanisms) may be an artifact of the response measure selected. The non-linear relations between psychophysical transformations (Gulliksen & Messick, 1960) have been virtually ignored in comparative surveys of taste interaction evidence, though they have been studied for a unidimensional taste (Ekman, 1961). Until these are more clearly understood, neither the existence nor the form of taste dimension interaction can safely be assumed.

ACKNOWLEDGEMENTS

I am very grateful to Professor G. M. Peterson of the Department of Mathematics, University of Canterbury, for comments and guidance; to Professor T. Indow of Keio University, Tokyo, for sending me unpublished data on taste interactions; and to Mrs R. M. Pangborn of the University of California, Davis, for supplying me with the data showing that earlier studies in this area had yielded results which were not subsequently replicated. The D^{**} computer program was written by Miss J. E. McKenzie of the Department of Psychology, University of Canterbury. Publication of this data was supported in part by funds from the University Grants Committee of New Zealand.

REFERENCES

AMERINE, M. A., PANGBORN, R. M. & ROESSLER, E. B. (1965). *Principles of Sensory Evaluation of Food.* New York: Academic Press.

BEEBE-CENTER, J. G., ROGERS, M. S., ATKINSON, W. M. & O'CONNELL, D. N. (1959). Sweetness and saltiness of compound solutions of sucrose and sodium chloride as a function of the concentration of solutes. *J. exp. Psychol.* **57,** 231–234.

BEEBE-CENTER, J. G. & WADDELL, D. (1948). A general psychological scale of taste. *J. Psychol.* **26,** 517–524.

BÉKÉSY, G. VON (1964). Duplexity theory of taste. *Science, N.Y.* **145,** 834–835.

BUJAS, Z. (1934). Quelques remarques sur le contraste et l'inhibition à la suite d'excitations gustatives simultanées. *C.R. Soc. biol.* **166,** 1304–1306.

DOVE, W. F. (1953). A universal gustometric scale in *D*-units. *Food Res.* **18,** 427–453.

EKMAN, G. (1961). Methodological note on scales of gustatory intensity. *Scand. J. Psychol.* **2,** 185–190.

EKMAN, G. & ENGEN, T. (1962). Multidimensional ratio scaling and multidimensional similarity in olfactory perception. (Rep. Psychol. Lab. Univ. Stockholm, No. 126.)

FABIAN, F. W. & BLUM, H. B. (1943). Relative taste potency of some basic food constituents and their competitive and compensatory action. *Food Res.* **8,** 179–193.

FISHER, R. A. (1934). Probability, likelihood, and quantity of information in the logic of uncertain inference. *Proc. R. Soc.* (*A*) **146,** 1–8.

Freeman, H. (1963). *Introduction to Statistical Inference.* Reading: Addison-Wesley.

Gregson, R. A. M. (1965). Representation of taste mixture cross-modal matching in a Minkowski *r*-metric. *Austral. J. Psychol.* **17**, 195–204.

Gregson, R. A. M. (1966 *a*). Qualitative identification of the taste of weak sucrose–sodium chloride mixture stimuli. *Percept. Psychophys.* **1**, 154–156.

Gregson, R. A. M. (1966 *b*). Theoretical and empirical multidimensional scalings of taste mixture matchings. *Br. J. math. statist. Psychol.* **18**, 59–75.

Gregson, R. A. M. (1966 *c*). *Scaling Perceived Similarities Between Complex Tastes which include Ethanol.* (Univ. Canterbury, Dept. Psychol. and Sociol. Res. Project 9.)

Gregson, R. A. M. (1966 *d*). Cross-modal matching of histograms and four-component taste mixtures, with one component fixed, under two pacing conditions. *Percept. mot. Skills* **23**, 183–190.

Gregson, R. A. M. & McCowen, P. J. (1963). The relative perception of weak sucrose–citric acid mixtures. *J. Food Sci.* **28**, 371–378.

Gregson, R. A. M. & Paris, G. L. (1967). Intensity–volume interaction effects in gustatory perception. *Percept. Psychophys.* **2**, 483–487.

Gregson, R. A. M. & Russell, P. N. (1965 *a*). Psychophysical power law experiment value for sucrose intensity. *Percept. mot. Skills* **20**, 294.

Gregson, R. A. M. & Russell, P. N. (1965 *b*). *Problems and Results in the Scaling of Intermodal and Intramodal Complex Taste Similarities by D* Metrics.* (Univ. Canterbury, Dept. Psychol. and Sociol. Res. Project 8.)

Gulliksen, H. & Messick, S. (1960). *Psychological Scaling: Theory and Applications.* New York: Wiley.

Hunt, E. B. (1965). The evaluation of somewhat parallel models. Chap. 5 in F. Massarik & P. Ratoosh, *Mathematical Explorations in Behavioural Science.* Homewood, Ill.: Irwin.

Indow, T. (1967). An application of the τ-scale of taste: interaction among the four qualities of taste. (Personal communication.)

Kamen, J. M., Pilgrim, F. J., Gutman, N. J. & Kroll, B. J. (1961). Interaction of suprathreshold taste stimuli. *J. exp. Psychol.* **62**, 348–356.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling; a numerical method. *Psychometrika* **29**, 115–129.

McBurney, D. M. & Lucas, J. A. (1966). Gustatory cross-adaptation between salts. *Psychon. Sci.* **4**, 301–302.

Minsky, M. (1961). Steps toward artificial intelligence. *Proc. I.R.E.* **49**, 8–30.

Pangborn, R. M. (1960). Taste interrelationships. *Food Res.* **25**, 245–256.

Pangborn, R. M. (1962). Taste interrelationships. III: Supra-threshold solutions of sucrose and sodium chloride. *J. Food Sci.* **27**, 495–500.

Russell, P. N. & Gregson, R. A. M. (1966). A comparison of intermodal and intramodal methods in the multi-dimensional scaling of 3-component taste mixtures. *Austral. J. Psychol.* **18**, 244–254.

Sebestyen, G. S. (1962). *Decision-making Processes in Pattern Recognition.* New York: Macmillan.

Shepard, R. N. (1962 *a*, *b*). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika* **27**, 125–139; 219–246.

Stevens, S. S. (1961). The Psychophysics of sensory function, In W. A. Rosenblith (ed.), *Sensory Communication.* New York: Wiley.

Wilde, D. J. (1964). *Optimum Seeking Methods.* Englewood Cliffs, N.J.: Prentice-Hall.

NOTES AND CORRESPONDENCE

NOTE ON A SIMPLE FIXED-POINT PROPERTY OF BINARY MIXTURES

By J. C. FALMAGNE[1]

F.N.R.S. University of Brussels, Belgium

Definition. A (binary) mixture of two distribution functions, H and K, is a function F defined for every real x by

$$F(x) = pH(x) + (1-p)K(x) \qquad (0 < p < 1).$$

Obviously, F is a distribution function. A simple fixed-point property of binary mixtures can be used as a test to decide whether a family of empirical distributions could be explained by a two-process model characterized as follows: any distribution in the family is the realization of a mixture of two constant distribution functions. Thus to each empirical distribution corresponds a value of the parameter p in the definition.

The following lemma is obvious.

Lemma. If H and K are defined as above, and if for some number x_0, $H(x_0) = K(x_0) = \alpha$ then any binary mixture of H and K has value α at x_0.

Theorem. Let F_1 and F_2, with parameters p_1 and p_2 respectively, be two mixtures of the distribution functions H and K. If $p_1 \neq p_2$, and if $F_1(x_0) = F_2(x_0) = \alpha$ for some x_0, then any mixture of H and K has value α at x_0.

For example, if F_1 and F_2 are continuous, and if their graphs cross each other at a point, the graph corresponding to any mixture of H and K must pass through the same point.

Proof. By the hypotheses of the theorem,

$$p_1 H(x_0) + (1-p_1)K(x_0) = p_2 H(x_0) + (1-p_2)K(x_0)$$

and thus

$$[H(x_0) - K(x_0)](p_1 - p_2) = 0.$$

We conclude that

$$H(x_0) = K(x_0) \quad \text{if} \quad p_1 \neq p_2.$$

The lemma establishes the theorem.

The interest of this theorem from an empirical standpoint is that, even though H and K are unknown, it enables conclusions to be drawn from an examination of the mixtures.

It is natural to raise a similar question concerning the mixtures of three or more distribution functions. As might be expected, more freedom is obtained in this case. For example, suppose that F_1 and F_2 are two mixtures of three basic distribution functions G, H and K, defined by

$$F_1(x) = p_1 G(x) + p_1' H(x) + (1 - p_1 - p_1')K(x),$$
$$F_2(x) = p_2 G(x) + p_2' H(x) + (1 - p_2 - p_2')K(x),$$

where the p's are strictly positive, and $p_i + p_i' < 1$. If, for some x_0, $F_1(x_0) = F_2(x_0)$, then the following equation can be derived

$$(p_1 - p_2)[G(x_0) - K(x_0)] + (p_1' - p_2')[H(x_0) - K(x_0)] = 0;$$

which does not seem to lead to any result of practical interest.

Finally, it should be noticed that the property expressed by the theorem holds for density functions and, more generally, for probability functions defined for an arbitrary space (with, in these two cases, the obvious definition of a mixture).

---

[1] At the University of Wisconsin during 1967–68.

S.P.

Example of an application

In Falmagne (1965) it was assumed that six empirical distributions of reaction time (RT) could be explained by a model involving two basic distributions and six values of the parameter p. The data shown in Fig. 1 correspond to Table 5 in that paper. It appears that the RT distributions are very different from each other. Consequently, the p values also are different. The theorem is thus applicable and shows that the model is untenable for any reasonable hypothesis about the basic distributions. For example, it can be seen that the graphs corresponding to the first two distributions must intersect at a point between 350 and 400 msec. The estimated value of the function at that point is indicated for each of the four remaining distributions. The fit to the last three is very poor. If the same general theoretical approach is to be retained, a mixture of not less than three distributions is necessary.

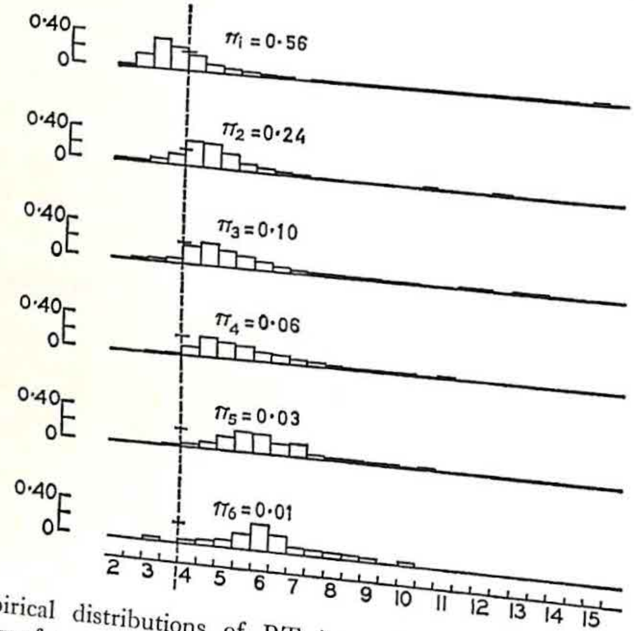


FIGURE 1. Empirical distributions of RT in a six-choice RT situation. π_i is the probability of presentation of stimulus i. The values on the abscissa represent the latency in 1/10th of a second. The dotted line indicates the latency at which the top two frequency distributions intersect. The horizontal bars on the line show the predicted common ordinate at this latency.

REFERENCE

FALMAGNE, J. C. (1965). Stochastic model for choice reaction time with applications to experimental results. *J. math. Psychol.* **2**, 77–124.

BOOK REVIEWS

Logic of Statistical Inference. By IAN HACKING. Cambridge University Press. 1965. Pp. ix + 232. 40*s.*

Statistical inference is a subject in which there are almost as many distinct points of view as workers in the field. Professor Hacking introduces here a new contribution.

The first question to be asked about such a contribution is what exactly the author means by ' statistical inference '. I think it is fair to say that there are two main ways in which we may be influenced by statistical information. It may change either our opinions or our decisions. Of course, these two aspects are interrelated. But in the case of decision-making we have at least some reasonably objective criteria to guide us. It may sometimes be possible to show that one decision is clearly better than another, in the sense of leading to greater financial profit or other desirable ends. Professor Hacking makes it clear that in his book he is not concerned at all with decisions, but only with beliefs. This is a much more thorny question. How are we to know whether one set of beliefs is more correct or appropriate than another? If someone asks us, ' Who are you to tell me what to believe? ' how are we to answer him? What sort of arguments can we put forward which will persuade him that we are right? But if statistics is to help us to come to judgements at all there must be some way of proceeding from the data to our conclusions. And anyone who wishes to do more than just blindly follow statistical procedures will have to think about this question seriously.

It is interesting to know what a philosopher and logician will make of current views on statistical inference. The book takes into consideration all points of view and topics related to the subject. It deals with the definition of probability, sampling, the minimax rule, likelihood, hypothesis testing, significance levels, sequential tests, fiducial probability, estimation, Jeffrey's theory, and the subjective theory. The discussion is always tolerant. Professor Hacking makes every effort to understand other authors' theories. It is penetrating: where objections are made, they are fairly and clearly made. It is often amusing, and enlivened by historical information. The style is lucid, and there are not many technicalities, either philosophical or mathematical, which are likely to impede the reader. However, the book is written in rather a tentative and exploring spirit, and does not lead to any precise conclusions of immediate practical value. Occasionally, one may feel that Professor Hacking has allowed himself to be persuaded into a rather odd point of view, although such lapses are rare. Thus on pp. 48–50 he considers the following situation (admittedly an artificial one, but introduced for the sake of argument). One knows that in a large store there are only three different kinds of urn, which contain respectively (*a*) 99 green balls, 1 red, (*b*) 2 green, 98 red, (*c*) 1 green, 99 red. One is given a large number of urns to examine, and they are ' produced from different and independent manufactories ' so that ' there is no reason to suppose the characteristics of one urn or machine are indicative of those of any other '. Furthermore and very importantly, though Professor Hacking does not say so explicitly, let us suppose that whoever produces the urns for examination is not deliberately trying to catch one out. From each urn one is allowed to select (at random) one and only one ball, and then one has to guess which type (*a*), (*b*) or (*c*) the urn belongs to, with the sole object of guessing right as often as possible. Common sense would suggest some strategy such as guessing (*a*) when a green ball is drawn, (*b*) or (*c*) when a red ball is drawn. The minimax strategy is very different: one takes a (very nearly true) coin, tosses it, and guesses (*b*) if it comes heads, (*a*) if it is tails and the ball is green, (*c*) if it is tails and the ball is red. Professor Hacking appears to say that this policy is not silly in the long run, although he would not himself use it; and he regards it as better than the more obvious ' commonsense ' rule. But he would regard it as silly if only one urn was being examined. It is not at all clear why Professor Hacking reaches these conclusions, which I would have thought clearly showed the strange behaviour of the minimax rule.

Professor Hacking's special contribution is to introduce a relation 'e supports i at least as well as d supports h' (which he writes symbolically in the form $h/d \leqslant i/e$. Here e, i, d, h are 'propositions'. The relation has to obey certain axioms, such as that if h implies i, then $h/e \leqslant i/e$. He claims that such a qualitative relation avoids possible difficulties which might arise when we try to give a quantitative measure to degrees of support or degrees of belief. It has, however, the disadvantage that since it is only qualitative, it would seem to be usefully applicable only to a rather limited range of situations.

The reviewer's feeling is that this is an interesting idea, but that progress will be held up until the central problem of what 'beliefs', 'opinions' and 'probabilities' are is properly tackled. These words indicate something going on in the mind: what is this something? This is a psychological question as well as a philosophical one. It is an extremely difficult one, but I doubt if the difficulty can be side-stepped. Philosophers tend to appeal to criteria such as 'the ordinary English use of the word "belief"'. This is dangerous. Ordinary language is a powerful instrument, but very badly adapted to the precise examination of abstract ideas. It is rather like trying to see electrons through a microscope; try as one will, the electrons, or ideas, will keep moving about. For this reason at the present stage of knowledge the consideration of the practical consequences of statistical decisions, which are reasonably clearly identifiable, seems more likely to yield helpful results than arguments about what beliefs are appropriate on given data. But obviously the weighing of belief is an important, if teasing, problem, and Professor Hacking's book is an amusing and stimulating contribution.

CEDRIC A. B. SMITH

Mathematical Epistemology and Psychology. By EVERT W. BETH and JEAN PIAGET. (Translated from the French by W. Mays.) Dordrecht, Holland: D. Reidel. 1966. Pp. xxi + 326. 54 florins.

This book owes its origin to Beth's hostile reaction to Piaget's *Traité de Logique*. Characteristically, Piaget suggested that since their views were so divergent the only useful thing to do was to prepare a joint work. Beth first wrote an historical survey of philosophical thinking about mathematics; Piaget then read this survey, and wrote his section on the 'genetic' approach to mathematical epistemology; and finally the authors combined to write a brief chapter of general conclusions. The collaboration, however, does not seem to have been particularly close—Beth makes no mention of Piaget's work—and, in essence, all that is finally agreed upon is that logic and psychology are autonomous disciplines.

Beth, as a logician, is chiefly concerned with the nature of mathematical reasoning. He establishes that it often involves a phase in which the mathematician reasons about a particular object, and that this reasoning differs from syllogistic reasoning. The influence of Kant's idea that mathematical demonstrations consist of a series of intuitive constructive steps is then traced through to the development of Brouwer's *intuitionism*. Beth then briefly examines 'psychologism', the attempt to base logic upon psychology, and decisively rejects it. There follows a long chapter on the *logicist* tradition, from Leibniz to Russell, the influence of Aristotle's methodology of the deductive sciences, a semi-technical account of Gödel's theorems, Gentzen's notion of natural deduction, and Beth's own technique of semantic tableaux. The latter is introduced primarily to demonstrate one possible solution to the problem of characterizing mathematical reasoning. The logically naive reader may have some difficulty with this chapter.

After the foundations, Beth turns to more psychological matters: the processes which underly mathematical creation. But his discussion is mainly limited to the well-known beliefs of Poincaré and Hadamard, and he does not say very much of interest. He writes (p. 91): 'The phase of preparation . . . cannot lead to profitable unconscious work, unless conscious work during this period is carried out in a reasonably effective way.' This, alas, is a truism, since no independent definition of 'reasonably effective' is forthcoming. For

Beth, the key to understanding mathematical creativity would be the construction of a typology of mathematical thought. This may be so; but it is a cause for regret that the collaboration failed to advance one. The final chapters of Beth's section are about the role and nature of mathematical intuition, and the relation between ' thinking machines ' and mathematical thought.

For Piaget, the fundamental problem is to discover a psychological basis for abstract logic and pure mathematics. But, just as Beth rejects ' psychologism ', Piaget rejects the idea, notably Marbe's, of introducing a logical factor as part of a psychological explanation. He then argues, however, that although the two disciplines are independent, they should be co-ordinated within ' genetic ' epistemology, which traces the successive stages of mental development. He claims that intelligence is ultimately based upon the interiorization of actions originally performed on objects in the physical world, and which become co-ordinated into ' operational structures '. These ideas are central to Piaget's theorizing and are worth examining in more detail.

A key concept is that of a reversible operation: an operation which annuls or negates the effect of a previous operation. In the early stages of operational thought, Piaget discerns three types of structure: those where the reversible operation is inversion or negation, involving classes; those where the reversible operation is ' reciprocation ', involving serial orders; and those whose elementary forms are of a topological character. Piaget is impressed by the correspondence between these three types of psychological structure and the structures which Bourbaki has postulated as fundamental to mathematics. But the correspondence is by no means exact: the psychological structures are what Piaget calls ' groupings ' (groupements)—primitive group-like structures which contain a number of restrictions preventing free combinatoriality. The ability to use natural numbers owes its origin to a ' synthesis ' of one grouping involving classes with another involving serial orders. The attainment of hypothetico-deductive intelligence, however, involves not just a ' synthesis ', but an actual combination of structures so as to contain both types of reversible operation. An individual at this stage of development (11 to 12 years of age) is operationally at home within the propositional calculus. He now possesses the inverse and reciprocal operations which, in combination with operations of ' correlativity ' and ' identity ', make up the so-called INRC group. Of course, the adolescent cannot formulate these operations; they are only manifest in his behaviour.

But why is it necessary to postulate the INRC group? Any such group of operations presupposes that the adolescent already possesses the basic connectives of the propositional calculus, and their expression in disjunctive normal form. And it is not altogether clear what further psychological function is served by the group operations upon these expressions —this is especially so for the operation of ' correlativity '. Piaget appears to introduce certain operations merely because he is committed to group structures. It is also doubtful whether adolescents have the logical ability attributed to them by this model. Piaget writes (p. 181) that, ' faced with a complex causal situation the subject will ask . . . whether fact x implies fact y (which he himself will often express by two propositions which we shall call p and q and which he will link by the words " if (p) then (q) "). To verify it, he will look in this case to see whether or not there is a counter-example x and non-y, therefore $p.\bar{q}$. ' Wason has found, however, that intelligent undergraduates presented with $\bar{q}$ fail to realize that it is important to test whether p may also be the case. Moreover, although a complex causal situation may be described by ' if p then q ', it is unlikely that this expression is meant as an instance of material implication. The reviewer has found, conversely, that if undergraduates have to describe the truth-table of material implication in their own words, they do not say ' if p then q ' but rather ' there's never a p unless there's a q ', or ' there isn't a p and a not-q together '.

Beth argued that the concept of *acquired* self-evidence should replace the Aristotelian notion of intuitive self-evidence. Piaget further suggests that mathematical concepts depend upon abstraction from operations, though they may be accompanied by some form of imagery or symbolic representation. Such intuitions are *not* derived by abstracting the common characteristics from a class of objects, as would be the case in acquiring the concept of physical weight. They are derived ultimately from *actions* performed upon objects, which may be generalized and abstracted, and in this way pure mathematics is freed from all constraints imposed by the physical world. This process of abstraction is never complete: thought may always be further extended and differentiated. But since the child's elementary structures will condition the process to some extent, their correspondence with the structures of Bourbaki is no accident.

Despite numerous disavowals, Piaget readmits a form of ' psychologism ' by stating that ' formal logic is the formalization of a certain form of natural thought . . . that of the logician himself '. And he makes no comment on Beth's espousal of Mannoury's ' psycholinguistics ' which demonstrates the discrepancies between formal logic and everyday reasoning.

The epistemological consequences of Piaget's approach are spelt out in the final chapter of his section. Simple empiricism, Platonism, nominalism, and the logical empiricism of Carnap, are all rejected, since they are inconsistent with the findings of ' genetic ' psychology. Much of the argument, however, seems to rely upon a rather cavalier treatment of the development of language. But what is of absorbing interest throughout Piaget's section is not the epistemology as such, but the insight it gives into the way Piaget thinks about mental development.

The book is marred by a number of misprints, including the Joycean ' sillogistics '. It is a difficult book to read because the authors, though preoccupied with structure, have failed to integrate their separate sections. But, if one perseveres, it is also a thought-provoking book. W. Mays deserves our gratitude for his translation.

P. N. JOHNSON-LAIRD

Test Theory. By D. MAGNUSSON. Reading, Mass.: Addison-Wesley. 1967. Pp. x + 270. 64*s*.

Dr Magnusson aims to provide ' a theoretically and statistically coherent introduction ' to test theory. To construe test theory as coherent would indeed be a scholarly task; to give students a coherent introduction would be doubly difficult. The barriers to coherence are many, but they are not unique to Magnusson's approach. First, there are at least two kinds of test theory: chapters may be divided into those which deal with the relations between items within a test and those concerned with relations between tests. Second, over the past 60 years the middle ground—reliability—has become littered with a heterogeneous collection of concepts, indices, methods and assumptions. Third, the truly psychological content of the theory is barely perceptible among the miscellaneous technologies and pragmatics of item analysis, reliability estimation methods, scoring methods, types of validity, corrections for guessing, administration difficulties, scale transformations, norms, and factor loadings. Hence the present text is again, basically, the standard cross-product of Gulliksen, Guilford and Ghiselli.

On the other hand, Magnusson's version does have a special flavour. He gives more space than is usual to the foundation work of Guttman and Coombs, Frederick Lord gets a fairer share than he has been accustomed to for his work on true score theory, and a shade more than lip-service is paid to Cronbach and Gleser's decision-theoretic approach. The more recent ' generalizability theory ' (1963: this journal) of Cronbach, Rajaratnam and Gleser is also mentioned several times, although Magnusson adopts his own more intuitive and descriptive approach to the variance components model. Whilst these and other references would give the student a broad hint about recent work on psychological measurement, there is a failure to integrate these approaches with classical test theory; or, if that is impossible, to provide either a rudimentary taxonomy or a survey of current directions.

As a teaching text it is neither better nor worse than others available. It is a point of some moment that those authors who would claim to simplify the symbols and reduce the algebra may in consequence produce a book which is neither satisfying to the mathematically competent, nor yet sufficiently attractive to those who are less competent. Particularly doubtful in this respect is the deliberate neglect of the distinction between population parameters and sample statistics, which is one of the links between the old and the new test theory. Perhaps the time has passd when a student of psychological tests can be allowed to skim over the algebraic expressions. Yet, though Magnusson recognizes that mathematical assumptions and deviations have important behavioural and operational implications, he chooses to write in another style.

PHILIP LEVY

Elementary Statistical Theory for Behaviour Scientists. By J. W. COTTON. Reading, Mass. and London: Addison-Wesley, 1967. Pp. x + 86. 15*s*.

This book, which sets out to supplement an ordinary introductory statistical course, can be divided into a section on distribution functions and densities, one on inference and the estimation of parameters, and one on various practical applications of statistics. Of these the first is undoubtedly the best and the student who perseveres with this section will have a more mature notion of what a continuous density is than the common one which regards it as a kind of blurred histogram.

The section on inference skims through the usual material including random sampling and how to estimate σ^2 from s^2; but no explanations are given. The author seems to be hurrying to get to the section on the desirable properties of estimators, though the explanations in this section are not very good. The minimum expected squared error is said to be a desirable property, for example, simply because many statisticians think that cost goes up as the square of the error, and several properties, such as maximum likelihood, are merely named. Bayesian estimation receives attention and Bayes' theorem is indeed derived, but no hint is given as to how posterior distributions are actually obtained.

The final section deals mainly with testing the difference between two means. Apart from the basic formula for the standard error of the mean, nothing is derived and results, such as the variance of a sum or difference, are simply stated. There is no attempt to justify the method of estimating σ^2 used in the t test, and the formulae are just stated. No reasons, either, are given for why the t distribution has to be used when σ^2 is estimated, nor is its distribution described. Likelihood ratios of both classical and Bayesian kinds are mentioned but nothing is said about their distributions, so that the student would not know how to use these ratios in order to evaluate his confidence in a decision. The Bayesian use of credible intervals for a parameter is compared with that of confidence intervals, but while this section will make the student think again about the usefulness of hypothesis testing, it is only two pages long. In view of all this, the book is more profitably viewed, not as an introduction to theory, but as a beginner's dictionary of statistical terms.

J. D. VALENTINE

Fundamentals of Behavioural Statistics. By R. RUNYON and A. HABER. Reading, Mass. and London: Addison-Wesley, 1967. Pp. x + 304. 40*s*.

It is claimed that this text is at an optimal level of difficulty for one semester introductory courses, that results are derived when this can be done simply, and that the authors have not been afraid to introduce new material. In fact, the computational formula for the standard deviation is almost the only derivation given, and Sandler's A-statistic (equivalent to the t test) for related means the only innovation.

The first, descriptive, part of the book would be difficult to distinguish from many other introductions, and more space is taken than necessary, e.g. a fresh illustrative histogram is given for each type of scale. There is also a sudden and unreasoned transition from discrete to continuous distributions, while the z distribution is introduced with details of areas between various z-scores, when surely this would have come better after the binomial distribution, where a reason could have been given for its shape. Correlation and regression are introduced as descriptive measures, but the regression coefficient is given as $b=r(s_x/s_y)$, an approach which completely fails to reveal the nature of b, whose derivation should be independent of that of r. Even when opportunities occur for easy derivations, as for the factor $(1-r^2)$ in the formula for departures from regression, the authors mainly state results.

In the second section concerned with inference, probability is well treated and its rules illustrated by enumeration. For continuous variables, however, probabilities are merely stated to be areas under the frequency curves without further explanation. Hypothesis testing is introduced by comparing the behaviour of a coin whose bias is being tested with the binomial distribution, and all the techniques of hypothesis testing (types of errors, etc.) follow clearly from this. Slips in this section include the statement that the two-tailed test is like asking, ' What is the probability of a score this rare? ' rather than ' this far from the mean '; and a false statement of the central limit theorem, whose true description goes under the head of the law of large numbers.

From now on the book departs more and more from its ideals. It follows the traditional route from z to t tests, but gives no help with why σ^2 is estimated as it is, nor why the t distribution differs from the z. The analysis of variance is usefully introduced through a comparison with the t test. But in spite of a consideration of the concept of sums of squares, nothing is said about why the between groups sum of squares is calculated as it is, nor is it pointed out that under H_0 both the between and the within group sums of squares estimate σ^2. The student, then, will probably not know what he is really doing, as he easily might. The major slip in this section is the recommendation of non-independent t tests for comparing means.

The final section on non-parametric tests begins with such a full discussion of power that afterwards the authors' energy is only up to giving the steps in the calculation of χ^2, the Mann–Whitney, and the Wilcoxon tests, for which no sort of logic or background is given. In the reviewer's opinion, the book is too inaccurate, and misses too many opportunities for good exposition, to be very strongly recommended.

J. D. VALENTINE

COMPUTER PROGRAM ABSTRACTS

Project Essay Grade: A FORTRAN Program for Statistical Analysis of Prose. By

E. B. Page, G. A. Fisher and Mary Ann Fisher, Bureau of Educational Research, University of Connecticut, Storrs, Connecticut, 06268, U.S.A.

Description This program (PEG-I) was written to collect information about ordinary English text that has been key-punched on to cards by typists. The original purpose of the analysis was to simulate human judgements of essay quality. The computer judgements were generated from the output of PEG-I by means of a standard multiple-regression program. For essays of American secondary students, such computer ratings were indistinguishable from those of ' expert ' teachers of composition (Page, 1966; Daigon, 1966; Page & Paulus, 1968). In the simulation, PEG-I provides 30 items of information about each student essay, to serve as measures and predictors. These include such variables as length (of essay, average sentence, average word), proportion (of connectives, prepositions, common words, mis-spellings), and various other counts and measures. It may therefore be used for investigations of prose. However, PEG-I is not limited to the present dimensions. It is a fairly general algorithm and may be readily adapted for other purposes of statistical content analysis.

A subroutine (Marcotte, 1966) permits the detection in the text of any specified phrases.

Language and Configuration PEG-I was written in FORTRAN IV for use with an IBM 7040 having 32K core storage. It uses and recognizes English words up to 12 letters long, ignoring any remaining letters. These are organized into double-precision arrays having two six-character columns, which are also addressed as complex numbers. Most of the FORTRAN routines are mnemonic and modular. Short MAP routines are used for packing single letters into the double-precision registers, and for unpacking them.

Input, Options, Output Input text may be on either cards or magnetic tape. Capital letters are not recorded, but simple conventions (such as ., for semi-colon) permit recognition of punctuation marks and paragraphs. Most key-punched English text could be easily adapted to the format.

Output is provided in two forms: (1) Each sentence is *printed* out in word-array form, with the obtained frequencies in an accompanying integer array. Summary frequencies and proportions are printed at the end of each essay output. (2) Only these summary data are *punched* out together with the essay's identification number. The output cards may then be used directly with any standard statistical programs.

Computation Times An essay of about 400 words requires approximately 40 sec. for analysis on the 7040, including time for organizing the strings into word arrays, and performing the various list lookup and the arithmetic tasks (but excluding phrase lookup).

Availability The listing and source deck, including PEG-I dictionaries, may be obtained by writing to the senior author at the University of Connecticut. A detailed explanation of its use is given by Page & Paulus (1968).

References

Daigon, A. (1966). Computer grading of English composition. *English J.* **57**, 46–52.

Page, E. B. (1966). Grading essays by computer: Progress report. *Proc.* 1966 *Invitational Conf. on Testing Problems*, pp. 87–100. Princeton, N.J.: Educational Testing Service.

Page, E. B. & Paulus, D. H. (1968). *The Analysis of Essays by Computer.* Final Report to the U.S. Office of Education. Storrs, Conn.: Bureau of Educational Research.

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Psychology (about 512 crown quarto pages per volume), *The British Journal of Medical Psychology* (about 350 crown quarto pages per volume), and the *British Journal of Social and Clinical Psychology* (about 320 crown quarto pages per volume) appear in quarterly parts, the subscription price of each, payable in advance, being £4 10*s*. per volume post-free ($16.50 in the U.S.A.). Subscriptions to any of these journals may be sent to any bookseller, or subscription agent, or to the Cambridge University Press, Bentley House, 200 Euston Road, London, N.W.1. American enquiries should be addressed to the Cambridge University Press, American Branch, 32 East 57th Street, New York, N.Y. 10022.

The British Journal of Mathematical and Statistical Psychology is issued twice a year, in May and November, in volumes of about 288 pages. The subscription rate, payable in advance, is £3 10*s*. per volume post free ($13 in the U.S.A.). Subscriptions should be sent to the Society's Agents, Taylor and Francis, Ltd., Red Lion Court, Fleet Street, E.C.4; or to any bookseller.

The British Journal of Educational Psychology is issued three times a year, in February, June and November. The subscription is £2 a year post free, payable in advance, or £1 each number. The subscription price in the U.S.A. is $4.20. Subscriptions should be sent to the Manager, B.J.E.P. Department, Methuen and Co. Ltd., 36 Essex Street, Strand, London, W.C.2.

The series of *Monograph Supplements* issued in connexion with *The British Journal of Psychology* is now discontinued. Orders may be sent to any bookseller, or the Cambridge University Press, 200 Euston Road, London, N.W.1, from whom details of these publications may be obtained.

MANUSCRIPTS

for *The British Journal of Psychology* should be sent to
Professor W. Sluckin, Department of Psychology, The University, Leicester;

for *The British Journal of Medical Psychology* should be sent to
Dr J. Sandler, Room 101, Department of Psychiatry, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London S.E.5;

for *The British Journal of Social and Clinical Psychology* should be sent to
Miss Thelma Veness (Social Psychology Editor), Department of Psychology, Birkbeck College, Malet Street, London, W.C.1, *or* to
Dr R. R. Hetherington (Clinical Psychology Editor), 6 Abercromby Square, Liverpool 7;

for *The British Journal of Mathematical and Statistical Psychology* should be sent to
Professor R. J. Audley, Department of Psychology, University College London, Gower Street, W.C.1;

for *The British Journal of Educational Psychology* should be sent to
Professor J. D. Nisbet, Department of Education, University of Aberdeen, Old Aberdeen.

SEPARATES

Contributors receive on application fifty copies of their papers free. Additional copies of separates may be obtained in units of fifty copies.

SUBSCRIPTIONS TO THE BRITISH PSYCHOLOGICAL SOCIETY

The subscription for Fellows, Associates and Graduate Members is £8 10*s*. (£7 10*s*. if paid by Banker's Order). They are entitled to purchase any or all of the following Journals at the special rates of £1 10*s*. per Journal per annum: *The British Journal of Psychology*, *The British Journal of Medical Psychology*, *The British Journal of Social and Clinical Psychology*, *The British Journal of Mathematical and Statistical Psychology* and *The British Journal of Educational Psychology*.

Information on membership and activities may be obtained from the Secretary, The British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1.

16 DEC 1968

THE BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY

Volume 21, *Part* 1, 1968

CONTENTS

BRELSFORD, JR., J. W., SHIFFRIN, R. M. and ATKINSON, R. C.
Multiple reinforcement effects in short-term memory *page* 1

MORTON, JOHN
A retest of the response-bias explanation of the word-frequency effect 21

SMITH, PHILIP TWITCHELL
Cost, discriminability and response bias 35

PASK, GORDON and LEWIS, B. N.
The use of a null-point method to study the acquisition of simple and complex transformation skills 61

JÖRESKOG, K. G. and LAWLEY, D. N.
New methods in maximum likelihood factor analysis 85

EVERITT, B. S.
Moments of the statistics kappa and weighted kappa 97

MAXWELL, A. E. and PILLINER, A. E. G.
Deriving coefficients of reliability and agreement for ratings 105

GREGSON, R. A. M.
Simulating perceived similarities between taste mixtures having mutually interacting components 117

NOTES AND CORRESPONDENCE
Note on a simple fixed-point property of binary mixtures 131

BOOK REVIEWS 133

COMPUTER PROGRAM ABSTRACTS 139

Printed in Great Britain by
Taylor & Francis Ltd., Red Lion Court, Fleet Street, London E.C.4

21/3/

Volume 21, Part 2

November 1968

THE BRITISH JOURNAL OF MATHEMATICAL & STATISTICAL PSYCHOLOGY

EDITED BY R. J. AUDLEY

WITH THE ASSISTANCE OF CYRIL BURT

TAYLOR & FRANCIS LTD
RED LION COURT, FLEET STREET, LONDON, E.C.4

40*s*. *net* $7.45

Book Review Editor: A. R. JONCKHEERE

Abstracts Editor (Computer Programs): H. J. HALLWORTH

Editorial Board:

C. Banks
M. S. Bartlett
G. H. Bower
D. E. Broadbent
J. Brown
V. R. Cane
L. J. Cronbach
C. I. Howarth
A. R. Jonckheere
D. N. Lawley
R. D. Luce
S. Messick
E. A. Peel
L. S. Penrose
J. Fraser Roberts
W. Stephenson
A. Stuart
P. E. Vernon

The British Journal of Mathematical and Statistical Psychology is issued by the British Psychological Society. Details of subscriptions to this and other Journals of the Society are given inside the back cover.

Papers for publication and books for review should be sent to the Editor, Professor R. J. Audley, Department of Psychology, University College London, Gower Street, W.C.1.

Abstracts of computer programs should be sent to Mr Owen White, Department of Psychology, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London S.E.5. Articles on computers, computer languages and programming will also occasionally be published and should be addressed to Mr White.

PREPARATION OF PAPERS

Contributors are asked to send their papers in a form which is ready for submission to the printer: that is to say, the arrangement of the material should follow that observed in a current issue of this *Journal*. Each article should be headed with a brief abstract. Each section should have a short cross-heading, in capitals; and the whole should conclude with a summary embodying the main conclusions reached. Special attention should be paid to such details as correct spelling, grammar, capitalization, numbering, etc. In preparing manuscripts and in correcting proofs authors should conform, so far as possible, with the 'Recommendations to Authors' set out in *The Printing of Mathematics* by T. W. Chaundy, P. R. Barrett, and Charles Batey (Oxford University Press, 1954, ch. 2, pp. 21–73), or to those given in the paper on 'Setting Mathematics' (Arthur Phillips, *Monotype Recorder*, 60, iv, 1956).

Owing to the present conditions for printing, the cost of setting up tables and algebraic formulae which require hand composition is very much greater than the cost of machine composition. Contributors are therefore advised to arrange their material so that it can be printed as economically as possible. Often numerical results and algebraic formulae can, with a little simplification, be readily adapted for purposes of machine composition. Authors must ensure that formulae are clear, particularly when these contain subscripts.

Tables, photographs and figures which are costly to print should be used only when they save considerable space or are essential for effective presentation; and in no case should both a figure and table be used to convey the same information. Each table or figure should be numbered and on a separate sheet with its approximate position indicated in the typescript.

Drawings should be in Indian ink on heavy unruled paper. Lettering on graphs, etc. is best indicated in a tracing placed over the diagram. Diagrams are normally reduced in reproduction and allowance must be made for a corresponding reduction in the thickness of lines and the spaces between them. The numbered legend of a figure should be typewritten on a separate sheet of paper, not lettered on the drawing.

The object of the caption for a table or the legend for a figure should be to make the table or the figure comprehensible without reference to the text.

References in the text should be indicated by dates in parentheses after the authors' names and correspond to an alphabetical list at the end of the paper. The list must follow the layout illustrated in the three following examples:

ALLPORT, G. W. & VERNON, P. E. (1931). A test for personal values. *J. abnorm. soc. Psychol.* **26**, 231–248.

GALTON, F. (1883). *Inquiries into Human Faculty and its Development*. London: Macmillan.

RÉVÉSZ, G. (1951). Le problème du génie. *Année psychol.* **50**, 83–96.

World list abbreviations are used for journals. The Oxford Dictionary is taken as the authority on spelling. For other general guidance, authors are advised to consult *Suggestions to Authors* issued by the Standing Committee on Publications of the British Psychological Society and obtainable, price 5*s.* (U.S.A. $1.00) post free, from Cambridge University Press.

In future authors will receive one complete slip-proof of their papers, but no page proof. They will be expected to pay the cost of all corrections for which they themselves are responsible.

All manuscripts should be submitted in duplicate.

PROGRAM ABSTRACTS

An abstract of a computer program, normally of 300 to 500 words in length, should state the sequence of standard mathematical or statistical operations that are performed, the language in which the program is written, the approximate running time and should either give an address from which the program can be obtained (in tape or card form) or state where data may be sent for processing. Authors are referred to a current issue of the *Journal* for examples of an acceptable abstract. In most cases, the abstract should be accompanied by an actual program in the form of a deck of cards or a paper tape together with the appropriate listing information, also any other instructions which are necessary for the preparation of data or the running of the program.

THE BRITISH JOURNAL OF MATHEMATICAL & STATISTICAL PSYCHOLOGY

EDITED BY R. J. AUDLEY

WITH THE ASSISTANCE OF CYRIL BURT

VOLUME XX1

TAYLOR & FRANCIS LTD
RED LION COURT, FLEET STREET, LONDON, E.C.4

Vol. 21 Part 2 141–159 The British Journal of Mathematical and Statistical Psychology November 1968

SLOW LEARNING

By M. Frank Norman
University of Pennsylvania

Learning will be slow if the quantity that represents the state of learning changes by small steps or with small probability on each experimental trial. Approximations to the distribution of this quantity for both types of slow learning are obtained. These approximations are applicable to a wide variety of models.

1. Introduction

There are at least two ways that learning can be slow. An experimental trial may produce small changes in a subject's habits, or it may produce no change at all with large probability. In the former case we say that learning occurs *by small steps*; in the latter we say it occurs *with small probability*. For example, consider a two-choice experiment in which a subject's probability of making response A after n trials (i.e. on the $(n+1)$st trial) is determined by a real quantity X_n, A's 'habit strength', $n \geqslant 0$. On trial $n+1$ any one of k events can occur. The jth of these has probability $\phi_j(X_n)$ and, if it occurs, then $X_{n+1} - X_n = \Delta X_n = U_j(X_n)$. Learning by small steps occurs if the increments have the form $U_j(x) = \theta V_j(x)$ and θ is positive but small. Learning occurs with small probability if one of the events, say the first, has no effect on habit strength ($U_1(x) \equiv 0$), and the event probabilities have the form $\phi_j(x) = c\gamma_j(x)$ $(j \geqslant 2)$, and $\phi_1(x) = 1 - c \sum_{j\geqslant 2} \gamma_j(x)$, where c is positive but small.

We will consider the behaviour of the Markov process $\{X_n\}$ as parameters like θ and c become small. To see what we should expect, consider the above example of learning by small steps under the added assumption that none of the V_j or ϕ_j depend on x and $\operatorname{var}(X_0) = 0$. Then

$$X_n - E[X_n] = \sum_{m=0}^{n-1} (\Delta X_m - E[\Delta X_m]),$$

so that, letting $Y_m = \Delta X_m / \theta$,

$$Z_n = (X_n - E[X_n])/\theta^{1/2} = (\theta n)^{1/2} \sum_{m=0}^{n-1} (Y_m - E[Y_m])/n^{1/2}.$$

Now the random variables Y_m are independent and identically distributed $(P(Y_m = V_j) = \phi_j)$. Hence the central limit theorem implies that as $n \to \infty$ and $\theta \to 0$ in such a way that $n\theta \to t > 0$, the distribution of Z_n converges to the normal distribution with mean 0 and variance $t\sigma^2$, where $\sigma^2 = \operatorname{var}(Y_m)$. Similarly, if $k = 2$, $U_2(x) \equiv 1$, $\gamma_2(x) \equiv 1$ and $P(X_0 = 0) = 1$ in the example of learning with small probability, then the ΔX_m are independent with $P(\Delta X_m = 1) = c$ and $P(\Delta X_m = 0) = 1 - c$. Thus the Poisson convergence theorem implies that as $n \to \infty$ and $c \to 0$ in such a way that $nc \to t$, the distribution of $X_n = \sum_{j=0}^{n-1} \Delta X_j$ converges to the Poisson distribution with mean t.

In this paper we will obtain extensions of the central limit and Poisson theorems sufficiently general to cover most of the instances of learning by small steps and with small probability that arise in learning models. We have made a point of considering multidimensional as well as unidimensional X_n since such processes, which arise, for instance, in models for choices among several alternatives or in multi-process models, are already important in learning theory and seem destined to become even more so. In all cases we will find the asymptotic distribution P_t of X_n (suitably normalized) as $n \to \infty$, $r (= \theta \text{ or } c) \to 0$ and $nr \to t$. In applications P_{nr} can then be considered an approximation to the distribution $F_n{}^r$ of X_n when the learning rate parameter r is small. Indeed, it follows from our results that, for any $T > 0$,

$$\max_{n:\, nr \leqslant T} d(F_n{}^r, P_{nr}) \to 0$$

as $r \to 0$, where $d(G, H)$ is a suitable measure of the distance between the distributions G and H. The restriction $nr \leqslant T$ on the natural time variable nr limits the approximation to the transient phase of a learning experiment.

Applications of the general theorems of Sections 2, 3 and 6 to some standard learning models will be given in Sections 4 and 7.

2. SMALL STEPS

2.1. *Assumptions*

Our assumptions in this section are abstracted from the example of learning by small steps given above. Let J be a bounded set of positive real numbers having 0 as a limit point. For every $\theta \in J$, let $\{X_n{}^\theta\}_{n \geqslant 0}$ be a Markov process with stationary transition probabilities in a closed (but not necessarily bounded) real interval I. We assume (a) that the conditional distribution of $Y_n{}^\theta = \Delta X_n{}^\theta / \theta$ given $X_n{}^\theta = x$ does not depend on θ. Let

$$w(x) = E[Y_n{}^\theta | X_n{}^\theta = x], \quad z(x) = E[(Y_n{}^\theta)^2 | X_n{}^\theta = x],$$
$$s(x) = z(x) - w^2(x) \quad \text{and} \quad r(x) = E[|Y_n{}^\theta|^3 | X_n{}^\theta = x].$$

We also assume (b) that $w(x)$ has two bounded derivatives and $s(x)$ has one, and (c) that $r(x)$ is bounded.

Under these assumptions we will determine functions $f(t, x)$ and $g(t, x)$ such that, if the initial state of $\{X_n{}^\theta\}$ is x ($P(X_0{}^\theta = x) = 1$), then $X_n{}^\theta$ is asymptotically normally distributed with mean $f(t, x)$ and variance $\theta g(t, x)$ as $n \to \infty$, $\theta \to 0$, and $n\theta \to t$ (Theorem 2.3). Since the variance of $X_n{}^\theta$ is small when θ is small (Corollary 2.2.2) the 'mean learning curve' $E[X_n{}^\theta]$ is especially significant. Our approximation $f(n\theta, x)$ to this quantity is relatively easy to compute. $f(\cdot, x)$ is the solution of the differential equation $f'(t) = w(f(t))$ satisfying the initial condition $f(0) = x$ (Lemma 2.2.3).

The picture we derive of a fundamental theoretical variable $X_n{}^\theta$ with a tightly clustered nearly normal distribution is reminiscent of the situation postulated by Hull (1943, p. 319).

2.2. *Two Lemmas*

Since $|w(x)| \leqslant z^{1/2}(x) \leqslant r^{1/3}(x)$, (c) implies that w, z and s are bounded. Let

$$a=\sup_{x\in I} s(x), \quad b=\sup_{x\in I}|w(x)|, \quad c=\sup_{x\in I}|w''(x)|,$$
$$d=\sup_{x\in I} w'(x) \quad \text{and} \quad g=\sup_{x\in I}|w'(x)|.$$

When x is the initial state of the process $\{X_n{}^\theta\}$, we sometimes denote the corresponding probabilities, expectations and variances by P_x, E_x and var_x, though, more often, especially in proofs, our notation will reflect neither x nor θ. Let

$$\mu_n=\mu_n(\theta, x)=E_x[X_n{}^\theta]$$

and

$$\omega_n=\omega_n(\theta, x)=\text{var}_x(X_n{}^\theta).$$

Lemma 2.2.1. For any $x\in I$, $n\geqslant 0$ and $\theta\in J$ sufficiently small that $\theta(2d+\theta g^2)\geqslant -1$,

$$\omega_n \leqslant \theta a A(n, \theta, 2d+\theta g^2),$$

where

$$A(n, \theta, h)=\theta\sum_{j=0}^{n-1}(1+\theta h)^j.$$

Proof.

$$\begin{aligned}\omega_{n+1} &= \text{var}(X_n+\Delta X_n)\\ &= \omega_n+\text{var}(\Delta X_n)+2\text{cov}(X_n, \Delta X_n).\end{aligned} \tag{2.2.1}$$

Now

$$\begin{aligned}\text{cov}(X_n, \Delta X_n) &= E[E[(X_n-\mu_n)(\Delta X_n-\Delta\mu_n)|X_n]]\\ &= E[(X_n-\mu_n)E[\Delta X_n-\Delta\mu_n|X_n]],\end{aligned}$$

and

$$\begin{aligned}E[\Delta X_n-\Delta\mu_n|X_n] &= \theta(w(X_n)-E[w(X_n)])\\ &= \theta((w(X_n)-w(\mu_n))+(w(\mu_n)-E[w(X_n)])),\end{aligned}$$

so that

$$\begin{aligned}\text{cov}(X_n, \Delta X_n) &= \theta E[(X_n-\mu_n)(w(X_n)-w(\mu_n))]\\ &= \theta E\left[(X_n-\mu_n)^2\frac{(w(X_n)-w(\mu_n))}{X_n-\mu_n}\right]\\ &\leqslant \theta d E[(X_n-\mu_n)^2]=\theta d\omega_n\end{aligned} \tag{2.2.2}$$

by the mean value theorem. Also,

$$\begin{aligned}\operatorname{var}(\Delta X_n) &\leqslant E[(\Delta X_n - \theta w(\mu_n))^2] \\ &= \theta^2 E[E[(Y_n - w(\mu_n))^2 | X_n]],\end{aligned}$$

and

$$\begin{aligned}E[(Y_n - w(\mu_n))^2 | X_n] &= s(X_n) + (w(X_n) - w(\mu_n))^2, \\ &\leqslant a + g^2(X_n - \mu_n)^2,\end{aligned}$$

so

$$\operatorname{var}(\Delta X_n) \leqslant \theta^2 a + \theta^2 g^2 \omega_n. \tag{2.2.3}$$

Combining eqns. (2.2.1) to (2.2.3) we obtain

$$\omega_{n+1} \leqslant (1 + \theta(2d + \theta g^2))\omega_n + \theta^2 a. \tag{2.2.4}$$

The conclusion of the lemma is obtained by iterating eqn. (2.2.4) and noting that $\omega_0 = 0$. Q.E.D.

If $h < 0$,

$$A(n, \theta, h) \leqslant 1/|h|,$$

whereas, for any h,

$$\begin{aligned}A(n, \theta, h) &\leqslant A(n, \theta, |h|) \\ &\leqslant [(1 + \theta|h|)^n - 1]/|h| \\ &\leqslant [\exp(n\theta|h|) - 1]/|h|.\end{aligned}$$

Thus we have the following corollary.

COROLLARY 2.2.2. (i) $\omega_n = O(\theta)$ uniformly in $n\theta \leqslant L$ for any $L < \infty$. (ii) If $d < 0$, $\omega_n = O(\theta)$ uniformly in $n \geqslant 0$.

By (i) we mean that there is a $\theta_0 > 0$ (in this case sup J is suitable) such that for any L there is a $B_L < \infty$ such that $\omega_n \leqslant B_L \theta$ whenever $\theta \leqslant \theta_0$ and $n\theta \leqslant L$. Similarly (ii) means that there is a $\theta_0 > 0$ (e.g. $|d|/g^2$) and a $B < \infty$ such that $\omega_n \leqslant B\theta$ for all $\theta \leqslant \theta_0$ and $n \geqslant 0$.

LEMMA 2.2.3. For any $x \in I$ there is a unique differentiable function $f(t) = f(t, x)$ such that

$$f'(t) = w(f(t)) \tag{2.2.5}$$

and

$$f(0) = x. \tag{2.2.6}$$

(i) $\mu_n - f(n\theta) = O(\theta)$ uniformly in $n\theta \leqslant L$ for any L.

(ii) If $d < 0$ then $\mu_n - f(n\theta) = O(\theta)$ uniformly in $n \geqslant 0$.

Proof.
$$\begin{aligned}\mu_{n+1} &= E[X_n+\Delta X_n]=\mu_n+E[\Delta X_n]\\ &=\mu_n+\theta E[w(X_n)]\\ &=\mu_n+\theta w(\mu_n)+\theta E[w(x_n)-w(\mu_n)].\end{aligned} \tag{2.2.7}$$

By Taylor's theorem

$$w(X_n)-w(\mu_n)=w'(\mu_n)(X_n-\mu_n)+\epsilon_n c(X_n-\mu_n)^2/2,$$

where here and in what follows ϵ_n denotes a quantity such that $|\epsilon_n|\leqslant 1$. Thus

$$E[w(X_n)-w(\mu_n)]=c\epsilon_n\omega_n/2.$$

Substituting this into eqn. (2.2.7) we obtain

$$\mu_{n+1}=\mu_n+\theta w(\mu_n)+\theta c\epsilon_n\omega_n/2. \tag{2.2.8}$$

Now if μ_n is to have an approximation of the form $f(n\theta)$, then $f(0)\doteq\mu_0=x$ and, by Corollary 2.2.2 and eqn. (2.2.8), $f((n+1)\theta)\doteq f(n\theta)+\theta w(f(n\theta))$, or $f'(t)\doteq w(f(t))$. Assume for the time being that I is the entire real line. Then standard theorems on ordinary differential equations (Birkhoff & Rota, 1962, Theorem 1, p. 103; Theorem 6, pp. 112–113) ensure that there is one and only one solution to eqns. (2.2.5) and (2.2.6). We will now prove that $\mu_n-f(n\theta)=O(\theta)$.

Let $\nu_n=\nu_n(\theta, x)=f(n\theta, x)$ and $n\theta=t$. Then

$$\nu_{n+1}-\nu_n-\theta w(\nu_n)=\int_0^\theta (f'(t+\tau)-f'(t))d\tau,$$

and

$$\begin{aligned}|f'(t+\tau)-f'(t)| &\leqslant g|f(t+\tau)-f(t)|\\ &=g|f'(t+\tau^*)|\tau\leqslant gb\tau,\end{aligned}$$

where $0<\tau^*<\tau$. Thus

$$|\nu_{n+1}-\nu_n-\theta w(\nu_n)|\leqslant gb\theta^2/2,$$

so that

$$\nu_{n+1}=\nu_n+\theta w(\nu_n)+\epsilon_n gb\theta^2/2. \tag{2.2.9}$$

Let $\Delta_n=\nu_n-\mu_n$. Subtracting eqn. (2.2.8) from eqn. (2.2.9) yields

$$\Delta_{n+1}=\Delta_n+\theta(w(\nu_n)-w(\mu_n))+\theta^2 r_n, \tag{2.2.10}$$

where

$$|r_n|\leqslant(gb+c\omega_n/\theta)/2.$$

Thus

$$\Delta_{n+1}=\left[1+\theta\frac{w(\nu_n)-w(\mu_n)}{\nu_n-\mu_n}\right]\Delta_n+\theta^2 r_n.$$

If $\theta g < 1$, the coefficient of Δ_n is positive, so that

$$|\Delta_{n+1}| \leqslant \left[1 + \theta \frac{w(\nu_n) - w(\mu_n)}{\nu_n - \mu_n}\right] |\Delta_n| + \theta^2 |r_n|$$
$$\leqslant (1 + \theta d)|\Delta_n| + \theta^2 |r_n|.$$

Iterating this equation, and noting that $\Delta_0 = 0$, we obtain

$$|\Delta_n| \leqslant \theta^2 \sum_{j=0}^{n-1} (1+\theta d)^j |r_{n-j}|$$
$$\leqslant \frac{\theta}{2}\left[gbA(n, \theta, d) + c\theta \sum_{j=0}^{n-1} (1+\theta d)^j \omega_{n-j}|\theta\right]. \tag{2.2.11}$$

The conclusions of the lemma then follow from eqn. (2.2.11) and Corollary 2.2.2.

It remains only to consider the case $I \neq (-\infty, \infty)$. Extend w to the entire line without increasing its supremum and in such a way that the Lipschitz condition $|w(x) - w(y)| \leqslant g|x-y|$ is satisfied for all real x and y. For example, if $B = \inf I > -\infty$ ($C = \sup I < \infty$) we can take $w(x) = w(B)$ ($w(x) = w(C)$) for $x \leqslant B$ ($x \geqslant C$). Eqns. (2.2.5) and (2.2.6) have a unique solution for the extended function w, and all of the work in the preceding paragraph is valid if d is replaced by g. Corollary 2.2.2(i) and eqn. (2.2.11) (with d replaced by g) imply that $\mu_n \to f(t)$ as $n\theta \to t$. But $\mu_n \in I$, so $f(t) \in I$ for all $t \geqslant 0$. Thus f is the unique solution of eqns. (2.2.5) and (2.2.6) for the unextended function w, and the work of the preceding paragraph is applicable as it stands. Q.E.D.

2.3. *The Main Theorem on Small Steps*

Let

$$g(t, x) = \int_0^t s(f(u, x)) \exp\left[2\int_u^t w'(f(v, x))dv\right] du. \tag{2.3.1}$$

THEOREM 2.3. *For any $x \in I$, if x is the initial state of $\{X_n^\theta\}$ for all $\theta \in J$, then the distribution of $Z_n = (X_n^\theta - \mu_n(\theta, x))/\theta^{1/2}$ converges to the normal distribution $P_{t,x}$ with mean 0 and variance $g(t, x)$ as $n \to \infty$, $\theta \to 0$ and $n\theta \to t \geqslant 0$.*

Proof. In order to keep the length of the proof within bounds and the main ideas in view, we will omit most details. The order of magnitude computations omitted are similar to those in the preceding section.

First we obtain a difference–differential equation for the characteristic function $h_n(\lambda)$ of Z_n. Observe that

$$h_{n+1}(\lambda) = E[\exp(i\lambda Z_{n+1})] = E[\exp(i\lambda Z_n)\exp(i\lambda \Delta Z_n)]$$
$$= E[\exp(i\lambda Z_n) E[\exp(i\lambda \Delta Z_n)|X_n]]. \tag{2.3.2}$$

For real y

$$e^{iy} = 1 + iy - y^2/2 + O(|y|^3),$$

thus

$$E[\exp(i\lambda\Delta Z_n)|X_n] = 1 + i\lambda E[\Delta Z_n|X_n] - \lambda^2 E[(\Delta Z_n)^2|X_n]/2 + |\lambda|^3 O(E[|\Delta Z_n|^3|X_n]). \qquad (2.3.3)$$

Using Corollary 2.2.2 (i) and Lemma 2.2.3(i) repeatedly, we obtain

$$E[\Delta Z_n|X_n] = \theta w'(\nu_n) Z_n + O(\theta^{3/2}(1 + Z_n^2)) \qquad (2.3.4)$$

and

$$E[(\Delta Z_n)^2|X_n] = \theta s(\nu_n) + O(\theta^{3/2}(1 + Z_n^2)), \qquad (2.3.5)$$

where $\nu_n = f(n\theta)$. Here and subsequently all O's are uniform when $n\theta$ is bounded. Also

$$E[|\Delta Z_n|^3|X_n] = \theta^{3/2} E[|Y_n - E[Y_n]|^3|X_n] \leqslant \theta^{3/2}(E^{1/3}[|Y_n|^3|X_n] + |E[Y_n]|)^3$$

by Minkowski's inequality. But $E[Y_n] = E[w(X_n)]$ is in absolute value no greater than $\sup_x|w(x)|$, and $E[|Y_n|^3|X_n] = r(X_n)$ is bounded by assumption (c). (Here we use the full force of (c) for the first time.) Thus

$$E[|\Delta Z_n|^3|X_n] = O(\theta^{3/2}). \qquad (2.3.6)$$

Substituting eqns. (2.3.4) to (2.3.6) into eqn. (2.3.3), substituting the resulting expression for the conditional characteristic function of Z_n given X_n into eqn. (2.3.2), and using Corollary 2.2.2(i), we obtain the desired equation

$$h_{n+1}(\lambda) - h_n(\lambda) = \theta[w'(\nu_n)\lambda h_n'(\lambda) - s(\nu_n)(\lambda^2/2)h_n(\lambda)] + O(\theta^{3/2}). \qquad (2.3.7)$$

The error term is of the order of magnitude of $\theta^{3/2}$ uniformly over n and λ when $n\theta$ and λ are bounded.

The difference–differential equation (2.3.7) suggests the partial differential equation

$$\frac{\partial H}{\partial t}(t, \lambda) = w'(f(t))\lambda \frac{\partial H}{\partial \lambda}(t, \lambda) - s(f(t))\frac{\lambda^2}{2} H(t, \lambda). \qquad (2.3.8)$$

We will show that the solution of this equation for which $H(0, \lambda) = h_0(\lambda) = 1$ is the characteristic function

$$H(t, \lambda) = \exp[-g(t)\lambda^2/2] \qquad (2.3.9)$$

of the normal distribution with mean 0 and variance $g(t)$, and that $h_n(\lambda) \to H(t, \lambda)$ as $\theta \to 0$, $n \to \infty$ and $n\theta \to t$. The key to both objectives is the reduction of eqn. (2.3.8) to a family of ordinary differential equations. Let $B(t)$ be the solution of the differential equation

$$B'(t) = -w'(f(t))B(t) \qquad (2.3.10)$$

for which $B(0) = 1$. That is,

$$B(t) = \exp\left[-\int_0^t w'(f(v))dv\right].$$

Eqn. (2.3.8) implies that

$$\frac{dH}{dt}(t, \xi B(t)) = -s(f(t))\frac{(\xi B(t))^2}{2} H(t, \xi B(t)), \tag{2.3.11}$$

so that, since $H(0, \xi) = 1$,

$$H(t, \xi B(t)) = \exp\left[-\frac{1}{2}\int_0^t (\xi B(u))^2 s(f(u))du\right],$$

for all t and ξ. Choosing $\xi = \lambda B^{-1}(t)$ we finally obtain eqn. (2.3.9). (For a general discussion of this technique, see Courant & Hilbert, 1962, pp. 62–66.)

For any real ξ, let $\xi_n = \xi B(n\theta)$, $H_n = H(n\theta, \xi_n)$, and $h_n = h_n(\xi_n)$. Eqn. (2.3.11) suggests that

$$H_{n+1} - H_n = -\theta s(\nu_n)\xi_n{}^2 H_n/2 + O(\theta^2), \tag{2.3.12}$$

and this is easily established via Taylor's theorem.

Also, writing

$$h_{n+1} - h_n = (h_{n+1}(\xi_{n+1}) - h_{n+1}(\xi_n)) + (h_{n+1}(\xi_n) - h_n(\xi_n)),$$

and applying eqn. (2.3.10) to the first and eqn. (2.3.7) to the second term on the right, we obtain the analogous equation

$$h_{n+1} - h_n = -\theta s(\nu_n)\xi_n{}^2 h_n/2 + O(\theta^{3/2}) \tag{2.3.13}$$

for h_n. Subtracting eqn. (2.3.13) from eqn. (2.3.12) we obtain

$$H_{n+1} - h_{n+1} = (1 - \theta s(\nu_n)\xi_n{}^2/2)(H_n - h_n) + O(\theta^{3/2}),$$

which implies that $H_n - h_n = O(\theta^{1/2})$. But $H_n \to H(t, \xi B(t))$ and $h_n - h_n(\xi B(t)) \to 0$ as $n\theta \to t$, hence $h_n(\xi B(t)) \to H(t, \xi B(t))$ as $n \to \infty$, $\theta \to 0$ and $n\theta \to t$, for all ξ and t. A change of variables $\xi = \lambda B^{-1}(t)$ and an application of the continuity theorem complete the proof. Q.E.D.

Another method of solving the differential equation (2.3.8) sheds light on eqn. (2.3.1). Assuming that eqn. (2.3.9) is valid for some differentiable function g, and substituting into eqn. (2.3.8), we obtain the ordinary differential equation

$$g'(t) = 2w'(f(t))g(t) + s(f(t)), \tag{2.3.14}$$

to which eqn. (2.3.1) gives the solution.

Let $\psi(t, \cdot) = \psi(t, \cdot, x)$ be the normal density with mean 0 and variance $g(t)$. The partial differential equation (2.3.8) can be obtained from the equation

$$\frac{\partial\psi}{\partial t}(t,y) = \tfrac{1}{2}s(f(t))\frac{\partial^2\psi}{\partial y^2}(t, y) - w'(f(t))\frac{\partial}{\partial y}(y\psi(t, y))$$

by Fourier transformation and integration by parts. This is the Kolmogorov forward or Fokker–Planck equation (Rosenblatt, 1962, p. 137) for a continuous time Markov process $Z(t)$ that may be regarded as an approximation to the process $(X_n{}^\theta - \mu_n)/\theta^{1/2}$ (t corresponds to $n\theta$) when θ is small. The nature of this and comparable approximations for the processes discussed in Sections 3 and 6, will be considered in a subsequent paper.

3. GENERALIZATIONS

3.1. *Generalizations of Theorem* 2.3

As is shown in Section 4, Theorem 2.3 covers the linear model for two-choice experiments and the beta model for such experiments when it is formulated in a certain way. Various slight modifications of the hypotheses of Theorem 2.3 yield a new theorem with a much broader range of applicability. The modifications encompass (i) multidimensional X_n^θ, (ii) weak dependence on θ of the distribution of $\Delta X_n^\theta/\theta$ given X_n^θ, and (iii) state spaces I_θ of $\{X_n^\theta\}$ that vary slightly with θ. Both (ii) and (iii) arise, for instance, in M element fixed sample size stimulus-sampling models for two-choice experiments, where $\theta = 1/M$ and X_n^θ is the proportion of stimulus elements conditioned to one of the response alternatives after n trials.

Let J be a bounded set of positive real numbers of which 0 is a limit point. Let N be a positive integer. R_N is the set of N-tuples of real numbers, which we regard as column vectors. For every $\theta \in J$, let $\{X_n^\theta\}_{n\geqslant 0}$ be a Markov process with stationary transition probabilities in a subset I_θ of R_N. Let I be the smallest closed convex set including all I_θ $(\theta \in J)$. Let

$$Y_n^\theta = \Delta X_n^\theta/\theta, \quad w(x, \theta) = E[Y_n^\theta | X_n^\theta = x],$$
$$s(x, \theta) = E[(Y_n^\theta - w(x, \theta))(Y_n^\theta - w(x, \theta))^T | X_n^\theta = x]$$

and

$$r(x, \theta) = E[|Y_n^\theta|^3 | X_n^\theta = x],$$

where T indicates transposition, and $|y|^2 = y^T y$ for $y \in R_N$. Thus $w(x, \theta)$ is the conditional mean vector and $s(x, \theta)$ the conditional covariance matrix of the normalized increment Y_n^θ, given $X_n^\theta = x$. For any $N \times N$ matrix A, let

$$|A|^2 = \sum_{i,j=1}^{N} a_{i,j}^2.$$

We assume that I_θ approximates I as $\theta \to 0$, in the sense that, (a1) for any $x \in I$,

$$\lim_{\theta\to 0} \inf_{y \in I_\theta} |x - y| = 0.$$

Next we suppose that there are functions $w(x)$ and $s(x)$ on I that approximate $w(x, \theta)$ and $s(x, \theta)$ when θ is small, by which we mean that (a2)

$$\sup_{x\in I_\theta} |w(x, \theta) - w(x)| = O(\theta) \quad \text{and} \quad \sup_{x \in I_\theta} |s(x, \theta) - s(x)| \to 0$$

as $\theta \to 0$. The function w is assumed differentiable in the sense that (b1) there is an $N \times N$ matrix valued function $w'(x)$ on I such that

$$\lim_{\substack{y\to x\\ y \in I}} \frac{|w(y) - w(x) - w'(x)(y - x)|}{|y - x|} = 0,$$

for all $x \in I$. We assume that $w'(x)$ is bounded: (b2)

$$\sup_{x \in I} |w'(x)| < \infty,$$

and that $w'(x)$ and $s(x)$ satisfy Lipschitz conditions (b3)

$$\sup_{\substack{x, y \in I \\ x \neq y}} \frac{|w'(x) - w'(y)|}{|x-y|} < \infty \quad \text{and} \quad \sup_{\substack{x, y \in I \\ x \neq y}} \frac{|s(x) - s(y)|}{|x-y|} < \infty.$$

Finally, we suppose that $r(x, \theta)$ is bounded: (c)

$$\sup_{\theta \in J,\, x \in I_\theta} r(x, \theta) < \infty.$$

For $\theta \in J$ and $x \in I_\theta$, let $\mu_n(\theta, x) = E_x[X_n{}^\theta]$, and $\omega_n(\theta, x) = E_x[|X_n{}^\theta - \mu_n(\theta, x)|^2]$. Under the above conditions we have the following theorem.

THEOREM 3.1. (i) $\omega_n(\theta, x) = O(\theta)$ uniformly in $x \in I_\theta$ and $n\theta \leqslant L$, for any $L < \infty$.

(ii) For any $x \in I$, the differential equation

$$f'(t) = w(f(t))$$

has a unique solution $f(t) = f(t, x)$ for which $f(0) = x$. $f(t) \in I$ for all $t \geqslant 0$.

$$\mu_n(\theta, x) - f(n\theta, x) = O(\theta)$$

uniformly in $x \in I_\theta$ and $n\theta \leqslant L$.

(iii) For any $x \in I$ the matrix differential equation

$$B'(t) = -w'(f(t, x))^T B(t)$$

has a unique solution $B(t) = B(t, x)$ for which $B(0)$ is the identity matrix. $B(t)$ is non-singular for all $t \geqslant 0$. For every $\theta \in J$ let $\{X_n{}^\theta\}$ have initial state $x_\theta \in I_\theta$. If $x \in I$, and $x_\theta \to x$ as $\theta \to 0$, then the distribution of $(X_n{}^\theta - \mu_n(\theta, x_\theta))/\theta^{1/2}$ converges to the normal distribution with mean 0 and covariance matrix

$$g(t, x) = \int_0^t [B(u, x) B^{-1}(t, x)]^T s(f(u, x)) [B(u, x) B^{-1}(t, x)]\, du$$

as $n \to \infty$, $\theta \to 0$ and $n\theta \to t$.

In view of (ii), we can, of course, substitute $f(n\theta, x_\theta)$ for $\mu_n(\theta, x_\theta)$ in (iii). It can be shown that $f(t, x) - f(t, x') = O(|x - x'|)$ uniformly in $t \leqslant L$ for any $L < \infty$, so $f(n\theta, x)$ can also be so substituted if $|x_\theta - x| = o(\theta^{1/2})$.

A proof of Theorem 3.1 can be obtained by straightforward but rather tedious modification of that of Theorem 2.3. We shall not give it here. The additional complications in the proof of the more general theorem are due almost entirely to the more complicated dependence of $\{X_n{}^\theta\}$ on θ, rather than to multidimensionality.

3.2. *Generalization of Corollary* 2.2.2(ii) *and Lemma* 2.2.3(ii)

We begin by noting some implications of negativity of $d=\sup_{x\in I} w'(x)$, within the framework of the assumptions of Section 2.1. The function $w(x)$ is bounded, so $I=[a, b]$ must be bounded, and w has at most one zero in this interval. If $w(a)<0$ then

$$\mu_1(\theta, a)=a+\theta w(a)<a,$$

an impossibility. Thus $w(a)\geqslant 0$. Similarly $w(b)\leqslant 0$. Thus w has a unique zero ρ, and $w'(\rho)<0$. These properties of w and boundedness of I are sufficient to obtain $\omega_n(\theta, x)=O(\theta)$ and $\mu_n(\theta, x)-f(n\theta, x)=O(\theta)$ uniformly in n, even under more general assumptions about the dependence of $\{X_n^\theta\}$ on θ.

We now state our assumptions precisely. J, $\{X_n^\theta\}$, I_θ, I, Y_n^θ and $w(x, \theta)$ are as in the second paragraph of Section 3.1 with $N=1$. Let

$$z(x, \theta)=E[(Y_n^\theta)^2|X_n^\theta=x].$$

We assume that (a1) and the first condition of (a2) hold, that $w(x)$ has a bounded derivative that satisfies a Lipschitz condition, and that $z(x, \theta)$ is bounded. Finally, we assume that I is bounded, that $w(x)$ has a unique zero ρ, and that $w'(\rho)<0$. Under these conditions we have this result.

THEOREM 3.2. (i) $\operatorname{var}_x(X_n^\theta)=O(\theta)$ uniformly in $x\in I_\theta$ and $n\geqslant 0$. (ii) For any $x\in I$, there is a unique differentiable function $f(t)=f(t, x)$ such that $f'(t)=w(f(t))$ and $f(0)=x$. $E_x[X_n^\theta]-f(n\theta, x)=O(\theta)$ uniformly over $x\in I_\theta$ and $n\geqslant 0$.

This theorem can be proved by suitably refining the arguments in the proof of Corollary 2.2.2(ii) and Lemma 2.2.3(ii). We omit the details.

Theorem 3.2(i) yields immediately

$$\limsup_{n\to\infty} \operatorname{var}_x(X_n^\theta)=O(\theta).$$

By Lemma 5.1 below, $f(t, x)\to\rho$ as $t\to\infty$, so that (ii) implies

$$\limsup_{n\to\infty} |E_x[X_n^\theta]-\rho|=O(\theta).$$

The problem of the asymptotic distribution of $(X_n^\theta-\rho)/\theta^{1/2}$ as $\theta\to 0$ *after* $n\to\infty$, under hypotheses like those of Theorem 3.2, has been considered by Norman & Graham (1968) and Norman (1968). As one might suppose on the basis of Theorem 3.1 and Lemma 5.4 below, the limiting distribution is normal with mean 0 and variance $s(\rho)/2|w'(\rho)|$.

4. EXAMPLES

Consider a two-choice (A_1 or A_2) learning experiment in which either response can be reinforced on any trial, regardless of what response the subject makes. Let π_{ij} be the probability that A_j is reinforced if A_i is made, and let p_n denote a subject's A_1 response probability after n trials ($n=0, 1, \ldots$).

In the linear model (see Sternberg, 1963) $\Delta p_n = \theta_{i1}(1-p_n)$ if response A_i on trial $n+1$ is followed by reinforcement of A_1, while $\Delta p_n = -\theta_{i2}p_n$ if A_2 is reinforced after A_i. In the beta model (see Sternberg, 1963) $p_n = v_n/(1+v_n)$, where $v_{n+1} = \beta_{ij}v_n$ if A_i is made and A_j is reinforced ($\beta_{i1} \geqslant 1$, $\beta_{i2} \leqslant 1$). At the level of p_n, the corresponding transformation is

$$p_{n+1} = \beta_{ij}p_n/[(1-p_n)+\beta_{ij}p_n],$$

while at the level of $y_n = \ln v_n$ it is $\Delta y_n = \ln \beta_{ij}$. In the M element stimulus sampling model with fixed sample size ν (see Atkinson & Estes, 1963) p_n is the proportion of stimulus elements conditioned to A_1 after n trials. If the stimulus sample on trial $n+1$ contains k elements conditioned to A_1, then A_1 is made with probability k/ν. If conditioning is effective, all elements in the sample become conditioned to the reinforced response (those already so conditioned do not change), so $\Delta p_n = (\nu-k)/M$ if A_1 is reinforced, while $\Delta p_n = -k/M$ if A_2 is reinforced. If conditioning is ineffective, $\Delta p_n = 0$. Conditioning is effective with probability c_{ij} if response A_j is reinforced after response A_i is made.

The theorems of the previous sections are applicable to all of these models. In the linear model we introduce the parameter θ by letting $\theta_{ij} = \theta d_{ij}$ for constants $d_{ij} \geqslant 0$. Thus the vector

$$(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$$

approaches 0 along a line in four-dimensional space as θ approaches 0. Taking $I = [0, 1]$ and $J = (0, 1/\max d_{ij})$, it is easy to verify that $X_n{}^\theta = p_n$ satisfies the hypotheses of Theorem 2.3. Here, as in all of the subsequent examples, the normalized increment $\Delta X_n{}^\theta/\theta$ is bounded, so the boundedness condition on its conditional third absolute moment is trivially satisfied. Clearly

$$w(p) = d_{11}\pi_{11}p(1-p) + d_{21}\pi_{21}(1-p)^2 - d_{12}\pi_{12}p^2 - d_{22}\pi_{22}(1-p)p.$$

If we take $\beta_{ij} = \exp(b_{ij}\theta)$ ($b_{i1} \geqslant 0$, $b_{i2} \leqslant 0$, $0 < \theta < \delta$), then $\ln \beta_{ij} = \theta b_{ij}$, and the process $X_n{}^\theta = y_n$ in the beta model satisfies the hypotheses of Theorem 2.3 with $I = (-\infty, \infty)$. Theorem 3.1 is applicable to p_n. For this process $I = I_\theta = [0, 1]$ and

$$w(p) = p(1-p)[a_1p + a_2(1-p)],$$

where $a_i = \pi_{i1}b_{i1} + \pi_{i2}b_{i2}$. The function

$$E[(\Delta v_n)^2 | v_n = v]$$

of v is typically unbounded, so Theorem 3.1 does not apply to the process v_n, regardless of how the parameter θ is introduced.

The hypotheses of Theorem 3.1 are satisfied for $X_n{}^\theta = p_n$ in the stimulus sampling model if we take $\theta = 1/M$, $J = \{1/M : M \geqslant 1\}$, $I_{1/M} = \{0, 1/M, \ldots, 1\}$, and $I = [0, 1]$. In this case it can be shown that

$$w(p) = \pi_{11}c_{11}p(1-p)(\nu-1) + \pi_{21}c_{21}(1-p)[p+\nu(1-p)]$$
$$- \pi_{12}c_{12}p[(1-p)+\nu p] - \pi_{22}c_{22}(1-p)p(\nu-1).$$

In the linear model, positivity of π_{ij} and d_{ij} for $i \neq j$ is necessary and sufficient for $w(0) > 0$ and $w(1) < 0$. Since w is quadratic, it then has a unique zero ρ in $[0, 1]$, and $w'(\rho) < 0$, so Theorem 3.2 is applicable. The same observation, with d_{ij} replaced by c_{ij}, is pertinent to the stimulus sampling model. The process p_n for the beta model never satisfies the hypotheses of Theorem 3.2 since $w(p)$ always has at least two zeros in $[0, 1]$.

As an example of a model for which X_n^θ is multidimensional we consider Wyckoff's discrimination learning model (see Bush, 1965, pp. 172–175). This model is partially specified by Bush's Table 1. To complete the specification we replace all plus and minus signs in the table by functions of the form $\theta\zeta(v)$, where v is the variable being transformed, $\zeta(v) \geqslant 0$ in the case of a plus and $\zeta(v) \leqslant 0$ in case of a minus. The functions ζ corresponding to different plus and minus signs may differ. All ζ's must satisfy $0 \leqslant v + \theta\zeta(v) \leqslant 1$ for $0 \leqslant v \leqslant 1$ and $0 < \theta < \delta$ in order to keep v in the unit interval, and all should have two continuous derivatives. Then Theorem 3.1 applies with $N=4$ if we take $X_n = (x_n, y_n, z_n, u_n)^T$, $J = (0, \delta)$, and $I = I_\theta$ the closed unit cube in R_4.

5. The Function f

In order to facilitate use of the approximation $f(n\theta)$ to $E[X_n^\theta]$ we collect here a number of observations concerning the function f of Theorem 3.1(ii) in the one-dimensional case.

Lemma 5.1. Let $x \in I$, and $f(t) = f(t, x)$.

(i) If $w(x) = 0$, $f(t) = x$ for all $t \geqslant 0$.

(ii) If $w(x) > 0$, $f'(t) > 0$ for all $t \geqslant 0$. If w has no zeros above x, $\lim_{t\to\infty} f(t) = \infty$. If y is the smallest zero of w above x, $\lim_{t\to\infty} f(t) = y$.

(iii) If $w(x) < 0$, $f'(t) < 0$ for all $t \leqslant 0$. If w has no zeros below x, $\lim_{t\to\infty} f(t) = -\infty$. If y is the largest zero of w below x, $\lim_{t\to\infty} f(t) = y$.

(iv) If $w(x) \neq 0$, $f = F^{-1}$ where

$$F(z) = \int_x^z \frac{1}{w(u)}\, du. \tag{5.1}$$

Proof. For any real z we denote the function on $[0, \infty)$ that is identically equal to z by z^*. If $w(x) = 0$, then x^* solves $x^{*\prime}(t) = w(x^*(t))$ and $x^*(0) = x$. Since $f(t)$ is the only function with these properties, $f(t) = x^*(t) = x$ for all $t \geqslant 0$.

Suppose now that $f'(0) = w(x) > 0$. If $f'(t) = 0$ for some t, let $\tau > 0$ be the smallest such t. Then $w(f(\tau)) = 0$, so $f(\tau)^*$ solves eqn. (2.2.5). Since this solution agrees with $f(t)$ at $t = \tau$, $f(t) = f(\tau)$ for all $t \geqslant 0$ (Birkhoff & Rota, 1962, Theorem 5, p. 20). In particular, $x = f(\tau)$. However, $f'(t) > 0$ for $0 \leqslant t < \tau$, so it is clear that $x < f(\tau)$. This contradiction shows that $f'(t) > 0$ for all $t \geqslant 0$.

It follows that $w(f(t))>0$ for all $t \geqslant 0$. Since this function is continuous, $1/w(f(t))$ is also. Rewriting eqn. (2.2.5) in the form

$$f'(\tau)/w(f(\tau))=1,$$

and integrating both sides between 0 and t we obtain

$$\int_0^t \frac{f'(\tau)}{w(f(\tau))}\,d\tau=t.$$

Under the change of variables $u=f(\tau)$ this becomes

$$\int_x^{f(t)} \frac{1}{w(u)}\,du=t,$$

which is (iv).

If w has no zeros above x, then $w(u)>0$ for all $u \geqslant x$. If $f(t) \leqslant B < \infty$ for all $t \geqslant 0$, then $\infty > F(B) \geqslant F(f(t))=t$, an absurdity. Thus if w has no zeros above x, $f(t) \to \infty$ as $t \to \infty$.

If w has a zero above x, let y be the smallest. If $f(t)=y$ for some t, then $f'(t)=0$. So $f(t)<y$ for all t. Now $w(u)>0$ for all $x \leqslant u < y$. If $f(t) \leqslant y-\delta$ for some $\delta>0$ and all $t \geqslant 0$, then $\infty > F(y-\delta) \geqslant t$ for all $t \geqslant 0$, which is impossible. Thus $f(t) \to y$ as $t \to \infty$.

The case $w(x)<0$ is similar to the case $w(x)>0$ treated above. Q.E.D.

Eqn. (5.1) reduces the determination of f to a quadrature and an inversion. In the three cases in the previous section where we have displayed a formula for w, this function is a polynomial. Whenever w is a polynomial, the quadrature can be reduced to integration of terms of the form $(u-a)^k$ by partial fractions expansion of $1/w(u)$. For instance, if w is quadratic with distinct (real) roots, say $w(u)=K(u-\lambda)(u-\gamma)$, this technique yields

$$f(t)=\frac{\gamma(x-\lambda)e^{K\lambda t}-\lambda(x-\gamma)e^{K\gamma t}}{(x-\lambda)e^{K\lambda t}-(x-\gamma)e^{K\gamma t}}.$$

This result covers the most difficult case that arises in the linear and stimulus sampling models discussed in Section 4. Almost all of the results concerning $E[p_n]$ for these models available in the literature apply only when $w(p, \theta)$ (and thus $w(p)$) is a linear function of p. Except for these cases, $f(n\theta)$ is the only analytical approximation to $E[p_n]$ currently available. A special case of this approximation was proposed by Bush & Mosteller (1955, eqn. 8.64, p. 183) and applied to experimental data with great success in their Sections 13.6 to 13.8. Their derivation suggests that the error will be small if $\operatorname{var}(p_n)$ is small, but they give no conditions for small $\operatorname{var}(p_n)$.

It is not always possible to express f in terms of elementary functions, and even when it is the expression need not be illuminating. Some information about f can be obtained directly from eqn. (2.2.5). For example, differentiating both sides of this equation we obtain

$$f''(t)=w'(f(t))f'(t).$$

If $w(x) \neq 0$, then $f'(t) \neq 0$, so $f''(t)=0$ if and only if $w'(f(t))=0$.

Lemma 5.2 gives simple asymptotic estimates for the deviation of $f(t)$ from its asymptote. Part (i) can be obtained directly from eqn. (2.2.5), while (ii) is a corollary of Lemma 5.1(iv). The proofs are omitted.

LEMMA 5.2. Suppose that $w(x) \neq 0$, and $f(t) \to y \in I$ as $t \to \infty$. Then $w(y) = 0$ and $w'(y) \leqslant 0$.

(i) If $w'(y) < 0$, then

$$G(x) = \lim_{z \to y} \int_x^z \left[\frac{w(u) - w(y)}{u - y} - w'(y) \right] \frac{du}{w(u)}$$

exists if z approaches y from the direction of x, and

$$f(t) - y \sim (x - y) \exp[G(x) + w'(y)t]$$

as $t \to \infty$.

(ii) If $|w(u)| \sim |u - y|^{1+\delta}\Lambda$ for $\delta > 0$ as u approaches y from the direction of x, then

$$|f(t) - y| \sim (\delta \Lambda t)^{-1/\delta}$$

as $t \to \infty$.

The final results in this section concern $g(t)$.

LEMMA 5.3. (i) If $w(x) = 0$, then

$$g(t) = s(x)(e^{2w'(x)t} - 1)/2w'(x).$$

(ii) If $w(x) \neq 0$, $g(t) = \Gamma(f(t))$, where

$$\Gamma(z) = w^2(z) \int_x^z \frac{s(u)}{w^3(u)} \, du.$$

Part (i) (read $g(t) = s(x)t$ if $w'(x) = 0$) is obtained from eqn. (2.3.1) and Lemma 5.1(i) by straightforward computation. Part (ii) can be established by introducing f as a new variable of integration in both of the integrals in eqn. (2.3.1).

LEMMA 5.4. If $f(t) \to y \in I$ as $t \to \infty$, and if $w'(y) < 0$, then

$$\lim_{t \to \infty} g(t) = s(y)/2|w'(y)|.$$

We will not prove this here. However, we note that the form of the limit is easily obtained by assuming that the limit exists and that $g'(t) \to 0$, and letting $t \to \infty$ in eqn. (2.3.14).

6. SMALL PROBABILITY

Let $\{X_n\}_{n \geqslant 0}$ be a Markov process with stationary transition probabilities

$$K(x, A) = P(X_{n+1} \in A | X_n = x)$$

in an abstract space I. The higher transistion probabilities

$$K^{(j)}(x, A) = P(X_{n+j} \in A | X_n = x) \tag{6.1}$$

are given recursively by

$$K^{(j+1)}(x, A)=\int_I K(x, dy)K^{(j)}(y, A). \tag{6.2}$$

From eqn. (6.1) we see that $K^{(0)}(x, A)=\delta_x(A)$, where

$$\delta_x(A)=\begin{cases}1 & \text{if} \quad x\in A\\ 0 & \text{if} \quad x\in' A,\end{cases}$$

and $K^{(1)}(x, A)=K(x, A)$. For $0<c\leqslant 1$, let $\{X_n{}^c\}$ be any Markov process with transition probabilities

$$K_c(x, A)=cK(x, A)+(1-c)\,\delta_x(A). \tag{6.3}$$

In the next section we will show that the parameter c discussed in Section 1 is of this type.

In order to see what we should expect of

$$P(X_n{}^c\in A|X_0{}^c=x)=K_c{}^{(n)}(x, A)$$

when n is large and c is small, we consider an example of a Markov process with transition probabilities (6.3). If $\{W_j\}_{j\geqslant 1}$ is a sequence of random variables with $P(W_j=1)=c$ and $P(W_j=0)=1-c$ that are independent of each other and of $\{X_n\}$, and if $S(n)=\sum_{j=1}^{n} W_j$, then $X_{S(n)}$ is such a process. Let $\{N(t)\}_{t\geqslant 0}$ be a Poisson process with mean inter-event time 1 that is independent of $\{X_n\}$. As $n\to\infty$, $c\to 0$ and $nc\to t$, the distribution of $S(n)$ converges to that of $N(t)$. Hence we expect the distribution of $X_{S(n)}$ to converge to that of $X_{N(t)}$, i.e.

$$K_c{}^{(n)}(x, A)\to L_t(x, A), \tag{6.4}$$

where

$$L_t(x, A)=\sum_{j=0}^{\infty} K^{(j)}(x, A)e^{-t}t^j/j!.$$

The Markov process $\{X_{N(t)}\}_{t\geqslant 0}$ is of pseudo-Poisson type (Feller, 1966, pp. 312–314).

Theorem 6.1 is somewhat stronger than expression (6.4). For any real y, $[y]$ is the least integer greater than or equal to y.

THEOREM 6.1. *For all n, $c>0$,*

$$|K_c{}^{(n)}(x, A)-L_{nc}(x, A)|\leqslant \exp(c[nc]\zeta(nc))-1,$$

where $\zeta(y)=1-([y]-1)/2y$.

It is easy to show that $\frac{1}{2}<\zeta(y)\leqslant 1$ for all $y>0$, $\zeta(y)=1$ for $0<y\leqslant 1$, and $\zeta(y)\to\frac{1}{2}$ as $y\to\infty$.

Proof. The formula

$$K_c^{(n)}(x, A) = \sum_{j=0}^{n} K^{(j)}(x, A) b_j, \tag{6.5}$$

where

$$b_j = \binom{n}{j} c^j (1-c)^{n-j}$$

follows easily from eqn. (6.3) by induction. Let $b_j = 0$ if $j > n$, and let

$$p_j = e^{-nc}(nc)^j / j!.$$

Clearly

$$\begin{aligned} K_c^{(n)}(x, A) - L_{nc}(x, A) &= \sum_{j=0}^{\infty} K^{(j)}(x, A)(b_j - p_j) \\ &\leqslant \sum_{j:\, b_j > p_j} (b_j - p_j) = D \end{aligned}$$

and

$$L_{nc}(x, A) - K_c^{(n)}(x, A) \leqslant \sum_{j:\, b_j \leqslant p_j} (p_j - b_j) = E.$$

But

$$0 = \sum_{j=0}^{\infty} (b_j - p_j) = D - E,$$

so $D = E$, and

$$|K_c^{(n)}(x, A) - L_{nc}(x, A)| \leqslant D. \tag{6.6}$$

Let $a_j = b_j / p_j$. Then

$$\begin{aligned} D &= \sum_{j:\, a_j > 1} p_j (a_j - 1) \\ &\leqslant \max_{j:\, a_j > 1} (a_j - 1) \sum_{j:\, a_j > 1} p_j \\ &\leqslant \max_{0 \leqslant j \leqslant n} a_j - 1. \end{aligned} \tag{6.7}$$

A routine computation yields

$$a_j = \frac{(n)_j (1-c)^{n-j}}{n^j e^{-nc}},$$

where $(n)_j = n \cdot (n-1) \ldots (n-j+1)$, so

$$\frac{a_{j+1}}{a_j} = \frac{n-j}{n-nc}$$

for $j \leqslant n$. Thus $a_{j+1} > a_j$ if $j < nc$, $a_{j+1} = a_j$ if $j = nc$ and $a_{j+1} < a_j$ if $j > nc$. It follows that

$$\max_{0 \leqslant j \leqslant n} a_j = a_{[nc]}. \tag{6.8}$$

Writing

$$a_j = (1 - 1/n) \ldots (1 - (j-1)/n)(1-c)^{n-j} e^{nc},$$

and applying the inequality $e^{-x} \geqslant (1-x)$ to all terms but the last, we obtain

$$a_j \leqslant \exp[jc(1-(j-1)/2nc)]. \tag{6.9}$$

The proof is completed by taking $j=[nc]$ in eqn. (6.9) and combining the resulting inequality with eqns. (6.6) to (6.8). Q.E.D.

7. EXAMPLES

The parameter c can be introduced into a broad class of learning models as follows. Suppose that the variable X_n that describes the state of learning after n trials takes on values in an abstract space I. Any one of k events can occur on trial $n+1$. The jth of these effects the transformation

$$X_{n+1} = F_j(X_n) \tag{7.1}$$

in the state of learning, and occurs with probability $\phi_j(X_n)$ $(j=1, \ldots, k)$. Naturally $\sum_{j=1}^{k} \phi_j(x) \equiv 1$ and $F_j(x) \in I$ if $\phi_j(x) > 0$. We now modify this model to take account of the possibility that the effect ascribed to the jth event in eqn. (7.1) occurs only with probability $c_j \in [0, 1]$, while with probability $1-c_j$, $X_{n+1} = X_n$. Thus eqn. (7.1) is applicable with probability $c_j\phi_j(X_n) = \phi_j^*(X_n)$, and

$$X_{n+1} = F_0(X_n), \tag{7.2}$$

where F_0 is the identity operator, with probability

$$\phi_0^*(X_n) = \sum_{j=1}^{k} (1-c_j)\phi_j(X_n).$$

The possibility that some of the F_j $(j \geqslant 1)$ are identity operators is not excluded.

Suppose that $c = \max_j c_j \neq 0$, and let $d_j = c_j/c$. Then

$$c_j = cd_j \quad \text{and} \quad \max_j d_j = 1.$$

To study the behaviour of the model when the c_j are small, we fix the vector $(d_1, \ldots, d_k)$ and let $c \to 0$. Let $\{X_n^c\}$ be the sequence of states corresponding to the superscripted value of the parameter, let

$$K_c(x, A) = P(X_{n+1}^c \in A \mid X_n^c = x)$$

be the transition probabilities for this Markov process, let $X_n = X_n^1$, and let $K(x, A) = K_1(x, A)$. To justify this notation and show that the theory developed in Section 6 is applicable, we must show that K_c and K satisfy eqn. (6.3). Now

$$K_c(x, A) = \sum_{0 \leqslant j \leqslant k:\, F_j(x) \in A} \phi_j^*(x)$$

and

$$\phi_0^*(x) = c\sum_{j=1}^{k} (1-d_j)\phi_j(x) + (1-c),$$

so that

$$K_c(x, A) = c \sum_{1 \leqslant j \leqslant k:\ F_j(x) \in A} d_j \phi_j(x) + \left[c \sum_{j=1}^{k} (1 - d_j) \phi_j(x) + (1 - c) \right] \delta_x(A)$$

$$= cK(x, A) + (1 - c)\delta_x(A).$$

Thus, indeed, eqn. (6.3) holds. It follows that Theorem 6.1 and its corollary (6.4) apply to $\{X_n{}^c\}_{n=0}{}^{\infty}$.

The construction described above may be used to introduce the parameter c into the linear, beta and Wyckoff models of Section 4. Linear models involving such a parameter have been considered by Norman (1964) and Yellott (1965). The parameters c_j have already been introduced into the stimulus sampling model discussed in Section 4, and to complete the construction it remains only to write $c_j = cd_j$. The parameter c is a standard feature of stimulus sampling theory.

ACKNOWLEDGEMENT

This research was supported by the National Science Foundation under grant GP–7335.

REFERENCES

ATKINSON, R. C. & ESTES, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. II, pp. 121–268. New York: Wiley.

BIRKHOFF, G. & ROTA, G. (1962). *Ordinary Differential Equations*. New York: Ginn.

BUSH, R. R. (1965). Identification learning. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. III, pp. 161–203. New York: Wiley.

BUSH, R. R. & MOSTELLER, F. (1955). *Stochastic Models for Learning*. New York: Wiley.

COURANT, R. & HILBERT, D. (1962). *Methods of Mathematical Physics*, vol. II. New York: Interscience.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications*, vol. II. New York: Wiley.

HULL, C. L. (1943). *Principles of Behavior*. New York: Appleton-Century-Crofts.

NORMAN, M. F. (1964). Incremental learning on random trials. *J. math. Psychol.* **1**, 336–350.

NORMAN, M. F. (1968). Mathematical learning theory. In G. B. Dantzig & A. F. Veinott (eds.), *Mathematics of the Decision Sciences*. Providence, R.I.: American Mathematical Society. (In press.)

NORMAN, M. F. & GRAHAM, N. V. (1968). A central limit theorem for families of stochastic processes indexed by a small average step size parameter, and some applications to learning models. *Psychometrika* **33**. (In press.)

ROSENBLATT, M. (1962). *Random Processes*. New York: Oxford University Press.

STERNBERG, S. (1963). Stochastic learning theory. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. II, pp. 1–120. New York: Wiley.

YELLOTT, J. I., Jr. (1965). Some effects of noncontingent success in human probability learning. *Technical Report No.* 89. Stanford, Calif.: Institute for Mathematical Studies in the Social Sciences.

Vo. 21
Part 2
161–182

The British Journal of Mathematical and
Statistical Psychology

November
1968

LATENCY AND RELATIVE FREQUENCY OF RESPONSE IN PSYCHOPHYSICAL DISCRIMINATION

By A. R. Pike
University of Strathclyde

The relation between response latency and probability is discussed with particular reference to differences between mean times for correct and incorrect responses in difficult two-choice discriminations. The predictions of some finite Markov chain models for this situation are considered, and possible modifications of these simple systems which may bring about features of observed data are discussed. Some results, in the form of curves of the latency–probability relation, are presented from earlier work and new discrimination experiments. These data are discussed in relation to theoretical interpretations based on a particular stochastic model which describes psychophysical discrimination in terms of the accumulation of units contributing to the overt response.

1. Introduction

In a paper by Audley & Pike (1965) it was pointed out that the relation between the probability of making a response and the time to initiate it is of considerable importance for testing models of choice. There are, of course, several ways of controlling response probability in choice situations, and the particular method considered in both the earlier paper and this one is that of varying the difficulty of a discrimination task. Difficult tasks clearly lead to response probabilities close to a half, and easy ones to probabilities close to zero and unity, respectively, for incorrect and correct responses. The experimental situation which will receive most attention here is one in which the subject has to decide which of two stimuli is farthest in one direction along some quantitative dimension. A number of discriminations varying in difficulty provide the required set of different relative response frequencies and their associated latencies.

The relation has not been studied with the detail it deserves either in psychophysics generally or choice theory in particular, the reason for this being partly historical and partly a question of convenience. Psychophysics has been preoccupied with problems of sensory resolution and ‘ laws ’ relating responses to stimuli, and many theorists have found it simpler to discuss a choice model only in terms of the relative frequencies with which different alternatives are chosen. The statistical decision theory of psychophysics is similar in this respect, exceptions being Stone (1960), Laming (1962), Carterette *et al.* (1965), Sekuler (1965) and Wolfendale (1967). In this paper a more detailed investigation will be made of the way in which the relation between latency and frequency may help in the investigation of the kinds of processes which may underlie decisions made in difficult two-choice discrimination tasks. It is then likely that the processes involve events associated with indecision which may have a comparatively long duration and the theoretical emphasis is placed on these events.

2. THE LATENCY–PROBABILITY FUNCTION

The relation under consideration will be referred to as the Latency–Probability Function (LPF). A refinement may be obtained by plotting the mean latencies as a proportion of the latency for a response probability of 0·5, though this requires, in the empirical case, a sufficient number of P points for reasonably accurate interpolation at $P=0{\cdot}5$. Fig. 1 is an example of this form of the graph for some theoretical processes to be discussed below.

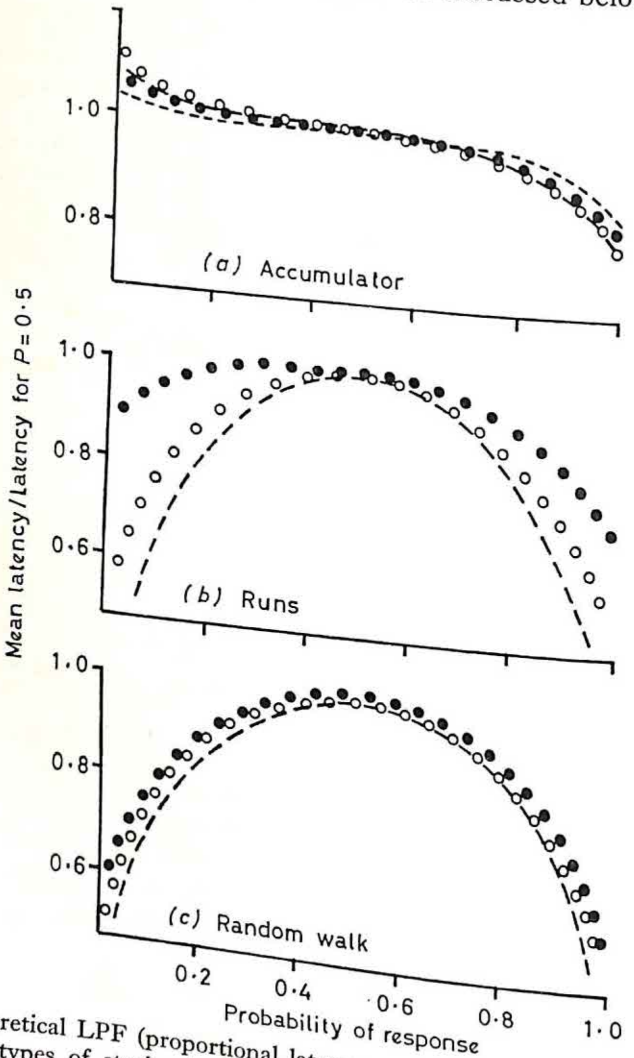


FIGURE 1. Theoretical LPF (proportional latency as a function of response probability) for three types of stochastic model for choice behaviour. ●, $k=2$; ○, $k=4$, — — —, $k=10$ (and - - - -, $k=50$, for accumulator).

It is clear that, in the two-choice situation, the probability of the dominant response (usually the correct response) will vary from 0·5 to 1 and the non-dominant response will co-vary from 0·5 to zero. Hence, for any range of discrimination difficulty or, in the theoretical case, the associated range of model parameters, pairs of LP points will be generated giving the required function. The general form of the graph will also give the difference in mean times between the correct (dominant) and incorrect (non-dominant) response for any particular probability of being correct. The term 'proportional time difference' will refer to such a difference derived from the proportional LPF.

Few theorists in the field of choice behaviour have been explicit about the form of the LPF, and not many more appear to have discussed the associated time difference between correct and incorrect responses in a two-choice task. The function was considered by Luce (1960), who derived the equation $L = a - b \log P$, and then proceeded to cast doubt upon this equation by reference to the data of Kellogg (1931) (Fig. 2). Luce points out that the mean latencies

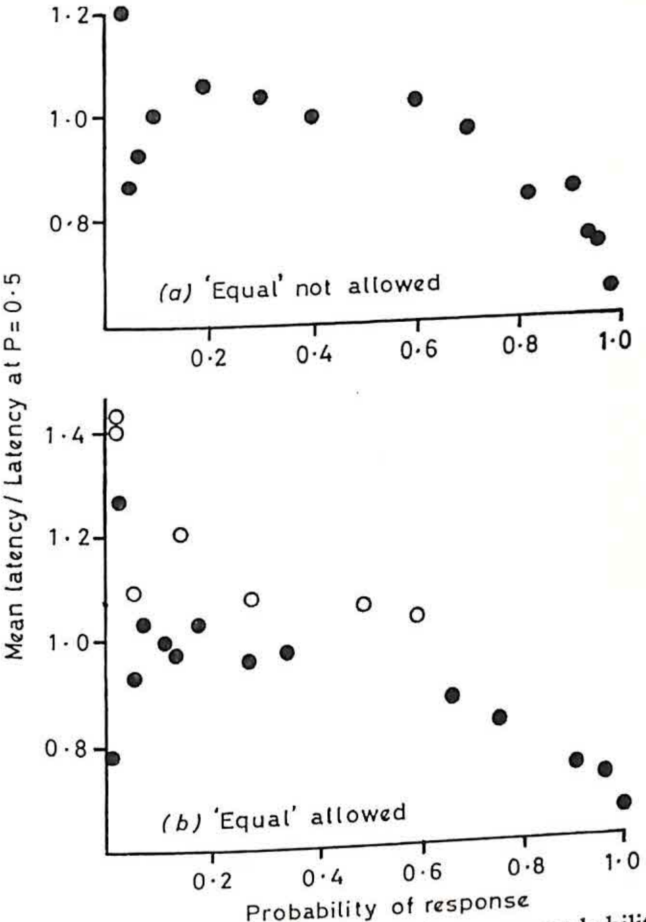


FIGURE 2. Proportional mean latency as a function of response probability in a judgement of light intensity. After Kellogg (1931). ○, response equal.

for $P < 0{\cdot}1$ are probably not very reliable: this is one of the chief difficulties in the use of the function and it arises, of course, because of the difficulty of obtaining sufficient data at low probabilities of response. It should be noted that Luce's equation will predict that correct mean latencies are shorter than those corresponding incorrect responses, which is a point in its favour, as will be seen later. Apart from Laberge (1962) and Audley & Pike (1965), there appears to be no other explicit study of the LPF. The 'recruitment' model of the former is for most purposes identical to the 'accumulator' model of the latter and will be discussed below: data given by Laberge (1962) in LP form are reproduced in Fig. 3.

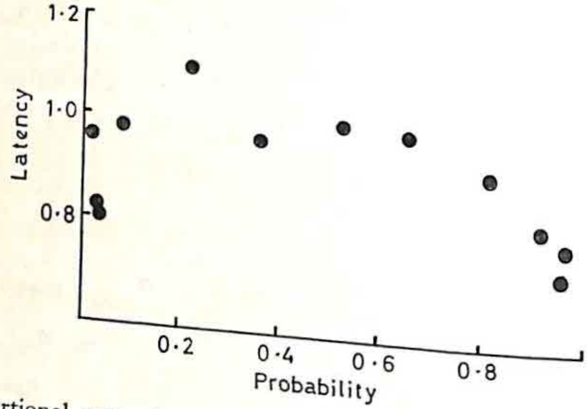


FIGURE 3. Proportional mean latency as a function of response probability: median latencies for 24 subjects in unbiased groups in a judgement of spatial position. After Laberge (1962).

Data giving both correct and incorrect mean times for the particular type of situation under discussion are also scarce. When these have been given for a range of discrimination difficulty and associated response probabilities it is possible to construct an LP graph. This has been done in Figs. 8 and 9 from the data of Wollen (1963) and Cross & Lane (1962) which will be considered below, together with other experimental results.

3. THEORETICAL INTERPRETATIONS OF LATENCY IN RELATION TO RESPONSE PROBABILITY

As mentioned above, few theories of choice have been explicit concerning the problem of differences between correct and incorrect latencies. It is, for example, difficult to obtain a definite prediction of the latency difference from the model of Cartwright & Festinger (1943), which is otherwise well developed (and 'fitted' to experimental data), and hence it will not be discussed. It appears that only the stochastic models and statistical decision models (e.g. Stone, 1960) have been stated explicitly enough to be compared with data obtained from psychophysical discriminations. (The latter models may usually be interpreted as a generalized form of ' random walk ' (cf. Feller, 1957) and may, therefore, be subsumed in the former.) Before dealing with the stochastic models, two theoretical approaches of a different kind will be briefly discussed.

3.1. *The Hull–Spence Approach*

The relation between response tendency and latency for the case of a single response has been discussed by Hull (1943) and Spence (1954) and it appears to have been interpreted by some as indicating an inverse relation in more complicated situations. Wollen (1963), for example, appears to assume that a weak response tendency will result in a longer latency in the two-choice case. However, such an assumption is by no means borne out if an attempt is made to interpret the Hull–Spence model in terms of competing responses. Spence deals only with relative frequencies in this situation but a derivation of latencies is possible. He shows that in the case of a single latency the postulate that Hull adopted ($\mathrm{RT} = aE^{-b}$, where E is effective excitatory potential) is inconsistent with earlier assumptions. He draws upon work of Estes (1950) showing that response latencies form a geometric distribution, resulting from the theoretical model describing the momentary oscillations of E about the fixed reaction threshold due to normally oscillating inhibition (i.e. there is a probability r that a response will not occur in any discrete time interval, given by the area of this normal distribution below the threshold value and hence a mean time to response of $1/1-r$). In dealing with competitive responses, Spence (1960) uses the same basic model, with each response having the same threshold and oscillation but different mean excitatory levels. It is possible to combine these ideas to form a latency model for the two-choice situation on a particular trial. According to the theory, in each moment of time from the beginning of the period when the organism is confronted with the choice-point there will exist (1) a probability that no response occurs, given by the sum of the areas of both oscillatory distributions below the threshold level; (2) a probability that the dominant response will occur, given by the product of the dominant area above and non-dominant area below plus the probability that if both are above the threshold the dominant has a greater excitatory level; and the remaining probability that the non-dominant response will occur. These probabilities may be derived from the values of the theoretical parameters and the results are given by Spence, but this is of no consequence here.

It is now easy to show that, for any specified constant values of the excitatory potentials—which would be the case if this model were to be adapted for the situation considered here—the resulting response distributions are geometric with identical mean response times. This is because in each moment of time one of three compound events, described above, will occur and the process continues to the first response event, giving a mean time of $1/1-r$, where r is the probability of no response in any time interval. Thus, in the competing situation the weak response tendency has the same mean time as its dominant counterpart, although it will have a longer mean time when considered singly. Further deductions may be made from certain changes in levels of excitatory potential during learning but they will not be considered here.

3.2. *A 'Measurement' Process*

One of the investigations into psychophysical discrimination which has discussed the role of the correct and incorrect latencies in relation to theoretical processes is that of Takada (1960), who studied the discrimination of the lengths of two lines. His theoretical model is an example of what may be termed a 'measurement' process, since the subject is assumed to perform some form of measurement operation during each discrimination, so that the accuracy of discrimination during a trial depends upon the efficiency of that process. The crucial point is that this efficiency is assumed to vary from trial to trial, with the more accurate 'measurements' taking a longer time to perform. Hence he predicts that correct mean times should be longer than incorrect ones, but finds that this is not the case in his own or other (Japanese) results. This kind of process may be given an interpretation in terms of information theory (so that the subject requires more information to bring about a correct response, hence requiring more time) and discrimination times have been discussed in the light of the formula given by Wiener (1961) for amount of information in a simple measurement task (i.e. $\log (a-b)^{-1}$, where a, b define the *a posteriori* region of indecision and 0, 1 the *a priori* region) by Welford (1960), who does not, however, discuss the time difference between correct and incorrect responses.

Because the prediction from this model, that correct mean times are longer, is not confirmed by the data presented in this paper it does not appear that a 'measurement' type of process will account for performance in difficult discriminations (i.e. where there is a probability of less than about 0·9 of being correct) and where the subject is not being instructed to respond at such a rate as to lower accuracy. However, it certainly cannot be ruled out for some discrimination, particularly when the discrimination itself is easy (e.g. see Laming, 1963); for in these cases there is little indecision in the system and if 'measurement' is at all possible, then it is, perhaps, likely that it will occur. It is not, of course, the only interpretation to explain longer correct times and Laming (1968), for example, uses the notion of variable times of commencement of information sampling to do so. Also, there is one version of a measurement process which is able to predict shorter times for correct responses in difficult tasks: Audley & Mercer (1968) have shown that the signal detectability model will do this by assuming that the closer the quality of information is to the criterion, then the longer the decision process for placement to one side or the other of that criterion. Hence, since the means of the incorrect portion of the 'information' (likelihood-ratio) distributions will usually be nearer the criterion than the correct means, the result follows. This is similar to the Coombs & Torgerson (Coombs, 1964, p. 530) interpretation in terms of the Thurstone distribution of differences between stimulus ' dispersions '. Shorter mean latencies for errors in choice-reaction tasks have been interpreted in terms of information capacity by Rabbitt (1966), S-R compatibility by Rabbitt & Phillips (1967) and in terms of ' perceptual focusing ' by Schouten & Bekker (1967).

Here, however, we are more interested in discrimination of a kind which is, by its very nature, more plausibly explained in terms of stochastic processes. The possibility that some form of indecision may occur in the system leads one to propose models in which that system assumes various pre-response states with particular probabilities, and the transitions between one pre-decisional state and another are likewise described stochastically. It is hoped that the arguments presented in this paper will make this proposal more acceptable. In the next section an attempt will be made to show that some fairly simple forms of stochastic process may account for the observed time differences and form of the LP graphs.

3.3. *The Stochastic Model for Choice*

Stochastic models for choice behaviour have been described by various writers, including Audley (1960), Estes (1960), Bower (1959), Laberge (1962) and Pike (1966). Usually, the processes proposed have been some simple form of finite Markov chain (cf. Feller, 1957); and the general term ' random walk ' has sometimes been used to describe them (e.g. Snell, 1965). Here a distinction will be made between three main types of model, and their predictions concerning response latencies in the two-choice situation will be examined. The division of the models into three main types follows Audley & Pike (1965) and little detail will therefore be given here. It will only be stated that each process consists of competing subsystems and proceeds from an initial zero state to a response (absorbing) state, that the threshold for a response is k pre-decisional events (steps, accumulations, run length) *of one kind* and that the probabilities associated with each pre-decisional event are taken to be p and q for the correct and incorrect response respectively, where $p+q=1$ and $p>q$. Discrete processes, with one such event occurring in each time interval, will be examined, but the comments made below generally remain unaltered for the continuous time processes (i.e. the events occur according to the Poisson process—see Feller, 1957).

The Accumulator Model

In this model the response threshold is defined as an accumulation of k pre-decisional events of one kind, regardless of the sequence of the different events in time. In the operation of this process, correct latencies are always shorter than incorrect latencies. This comes about for the following reason. The ratio of the (conditional) probability of a correct to that of an incorrect response at discrete time n decreases as n is increased from k to $2k-1$, the latency range of this process. This ratio is of the form

$$R_n=(p/q)^{2k-v} \quad (k \leqslant v \leqslant 2k-1)$$

for the response at $n=v$. Hence, the incorrect latency distribution will have a greater mean than the correct distribution. This latency difference increases with the value of k, but it appears that the LPF will remain fairly constant in form if expressed as in Fig. 1 *a*. An expression for the mean latency of correct responses is given by:

$$u_p = \frac{k}{p} - \frac{2k-1!}{k-1!k-1!} \; \frac{q^k p^{k-1}}{P},$$

where P is the response probability. This is derived, by differentiation, from the generating function presented in Audley & Pike (1965), and the mean latency for incorrect responses is similarly expressed.

The Runs Model

Here an overt response occurs when a run of k events of one kind first occurs. As may be seen from Fig. 1, the LP graph for this process varies somewhat for the lower values of k: for $k=2$ it is decidedly asymmetrical and correct latencies are shorter than incorrect ones. However, with increasing k the function quickly becomes symmetrical and the proportional time differences tend to zero. For example, the theoretical latencies for $p=0{\cdot}7$ are as follows:

k	*Correct (C)*	*Incorrect (I)*	*Proportional difference I-C/Mean*
2	2·75	2·96	0·075
4	14·15	14·35	0·014
10	930	930·5	0·00054

The asymmetry is again due to variation in the ratio R_n of conditional probabilities of response at time $k, k+1, \dots$ (an infinite range in this case), but this variation quickly becomes negligible as the threshold value is increased.

Random Walk (Simple One-dimensional) Model

In this case the process may be described in terms of a point moving to the right or left along a line with probabilities p and q until it reaches a point k steps away from the origin of the walk in either direction. The prediction from this process is a symmetrical LP curve for all values of k, as shown in Fig. 1. By considering the ratio R_n this conclusion follows immediately, since this ratio is constant at the value $(p/q)^k$ (cf. Feller, 1957). Hence, there is no latency difference between the two responses at any given probability. It is an interesting and useful fact that all the moments of the latency distribution will also display a similar symmetry in the unbiased (i.e. equal threshold) case. This is because the moment generating functions are of the form $p^k g_k(\theta)$ and $q^k g_k(\theta)$ (cf. Feller, 1957).

4. Factors affecting the Form of the LPF

Modifications may be made in the three stochastic processes which lead to a time difference between correct and incorrect responses whether or not it already exists in the basic models. The main forms of these will now be discussed.

4.1. *Bias*

It is apparent that any of these systems could be modified to represent phenomena corresponding with empirical bias towards one response or another. Such bias may be regarded as being due to sensory, perceptual or response factors. Bias which is judgemental in origin is also conceivable, and this would therefore bear some relation to variation in the decision criterion as in signal-detection theory. In stochastic systems of the above kind, bias may be described in terms of unequal values of the threshold for response (k-bias) or it may be associated with the values of p and q as, for example, when the left-hand member of a pair of stimuli always appears to be relatively greater. In the latter case, a bias parameter, either multiplicative or additive, would be included in the model. One consequence of these modifications is to bring about a relative decrease in latency for the response toward which the system is biased, although it is not generally possible to describe these effects in any simple manner. However, in the case of p-bias—and for threshold bias in the random walk—it is possible to derive the consequences simply in the following way.

Consider the general expression for the difference between correct and incorrect latencies when a position bias exists in the two-choice situation. This means that the response probabilities and latencies for correct and incorrect responses will differ according to the stimulus arrangement and must be averaged to obtain the overall values. Specifically, if P_1, P_2, u_1 and u_2 are the correct response probabilities and latencies and Q_1, Q_2, v_1, v_2 the respective incorrect probabilities and latencies, then the difference between incorrect and correct mean latencies (or between non-dominant and dominant latencies) is given by

$$\begin{aligned} u_q - u_p &= \frac{Q_1 v_1 + Q_2 v_2}{Q_1 + Q_2} - \frac{P_1 u_1 + P_2 u_2}{P_1 + P_2} \\ &= \frac{P_1 Q_1 (v_1 - u_1) + P_1 Q_2 (v_2 - u_1) + P_2 Q_1 (v_1 - u_2) + P_2 Q_2 (v_2 - u_2)}{(P_1 + P_2)(Q_1 + Q_2)}. \qquad (1) \end{aligned}$$

In the case of p-bias, k remains constant and the sign of the latency difference may be deduced from the known form of the unbiased LPF; assuming that $P_1 > P_2 > 0{\cdot}5$ so that the four points of the curve relevant to eqn. (1) are, from left to right, $(Q_1 v_1)$, $(Q_2 v_2)$, $(P_2 u_2)$, $(P_1 u_1)$, with the order of the L values depending upon the form of the function.

For the random walk, $u_1 = v_1$ and $u_2 = v_2$: also, since $P_1 > P_2$, $u_1 < u_2$ and $v_2 - u_1 = u_2 - v_1$. Hence, in eqn. (1) the first and last terms in the numerator vanish and the remainder is positive since the second term exceeds the third.

Thus this form of bias in the random walk produces incorrect latencies which are longer than correct latencies. An example is given in Fig. 4. It should be noted, however, that the average latencies for the different stimulus arrangements will remain equal since the values of k are equal.

In the runs process the assumption may be made that the LP curve has a maximum at $P=0{\cdot}5$ and the left side increases more slowly than the right side decreases, so that dominant latencies are less than non-dominant. There are now two cases of eqn. (1) to consider: the third term in the numerator is either negative or positive, since u_2 may be greater or less than v_1; but in each case the latency difference is positive, since term 2 will exceed term 3, and the same conclusion follows as for the random walk. For other than very small values of k the separate proportional time differences should be very nearly zero.

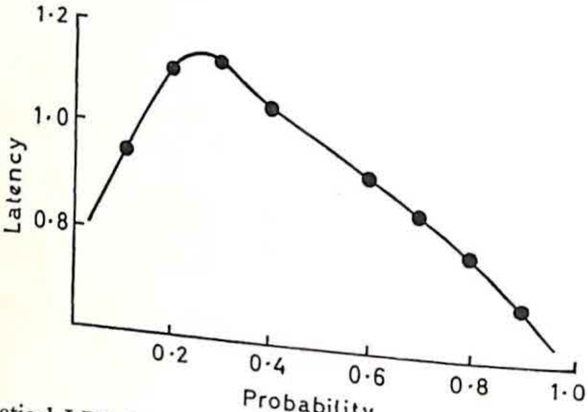


FIGURE 4. Theoretical LPF for a random walk process with additive bias ($p=p'\pm0{\cdot}1$).

In the case of the accumulator, all the terms of the equation are positive because of the continual decrease of the LPF: the overall effect is slight and an example is given in Fig. 5. In this model there is no requirement that the average latencies for different stimulus arrangements be equal.

Thus, when p-bias occurs in the stochastic processes considered here, all the theoretical LPF will show shorter latencies for the dominant response.

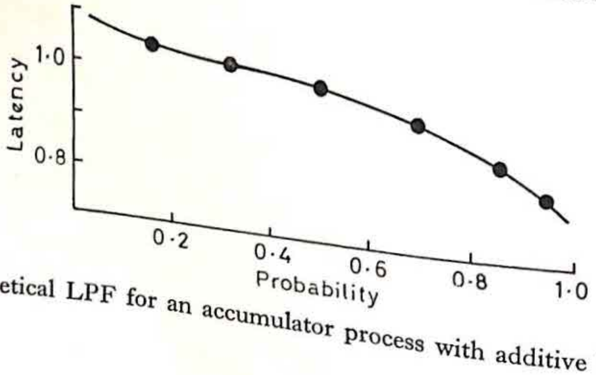


FIGURE 5. Theoretical LPF for an accumulator process with additive bias ($p=p'\pm0{\cdot}1$).

Conclusions concerning k-bias are arrived at easily only in the case of the random walk, because of the special properties of its generating function mentioned above. The differing values of k are now constant with respect to stimulus position, and the values of p and q are constant but alternate as the stimuli are changed. Hence, the response probabilities will change according to stimulus arrangement. The mean latencies are no longer equal but remain constant with respect to position, the differing probabilities producing an overall difference in the correct and incorrect latencies. In fact, all the moments (i.e. the derivatives of the m.g.f.s weighted by response probability) will show this feature. It follows from the fact that the m.g.f.s are now:

$$p^{k_1} u(\theta), \qquad q^{k_2} v(\theta)$$

for one stimulus arrangement (k_1 and dominant stimulus on left) which has the associated response probabilities

$$p^{k_1} u(0) = P_1, \qquad q^{k_2} v(0) = Q_1,$$

and

$$q^{k_1} u(0), \qquad p^{k_2} v(0)$$

for the other (k_1 and non-dominant stimulus on left) and with

$$q^{k_1} u(\theta) = Q_2, \qquad p^{k_2} v(\theta) = P_2.$$

The functions u, v differ only in the parameter k and the mean latencies are given by

$$u_1 = u'(0)/u(0) \qquad \text{and} \qquad u_2 = v'(0)/v(0)$$

for the responses associated with k_1 and k_2 respectively. It can be shown from the latencies expressed in terms of k and p, if it is not otherwise obvious, that when $k_1 < k_2$ then $P_1 > P_2$ and $u_1 < u_2$. Hence, since $u_1 = v_2$ and $u_2 = v_1$, the numerator of eqn. (1) reduces to

$$(u_2 - u_1)(P_1 Q_1 - P_2 Q_2),$$

which is always negative. Thus, it must be concluded that correct latencies are longer than incorrect latencies in this case.

This form of bias in the accumulator produces a flattening of the LPF, or so it appears from some computed examples, but this will not be illustrated. The runs process again changes its LPF, so that incorrect latencies are considerably lengthened. This would be particularly important for this model; it can now conform more readily with experimental data since the mean latencies of the separate response categories are now by no means approximately equal, varying according to k_1 and k_2. An example of this biased process LPF is given in Fig. 6.

Bias is, of course, an important factor in signal-detection studies and this will be referred to below.

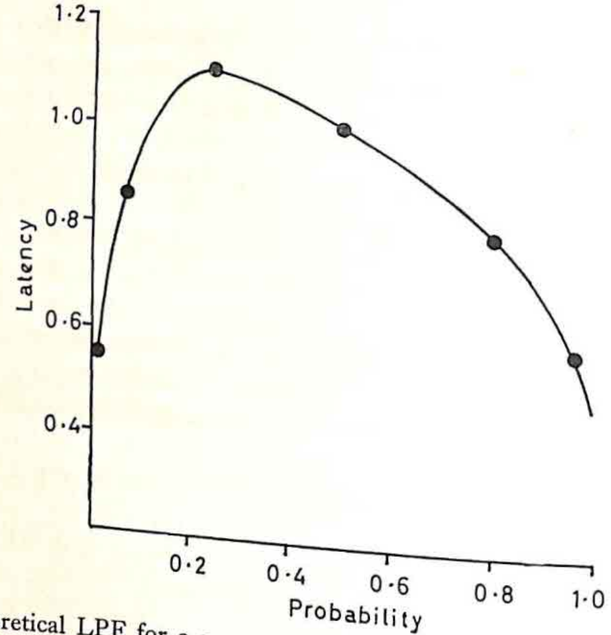


FIGURE 6. Theoretical LPF for a runs process with $k=4$ and 6 (i.e. a biased response threshold).

4.2. *Heterogeneity*

Suppose the values of the underlying parameters p, q can vary from trial to trial. It is clear that when the value of p for the correct response (necessarily $>0{\cdot}5$) is at a higher value than its average, the response time will be shorter and there is a greater probability of the correct response occurring (i.e. the competition in the process is more one-sided than at the average value of p). The opposite effect occurs when the lower values, nearer to 0·5, are in operation (and simultaneously, of course, the higher values for q). Hence the time difference is brought about for an averaged series of trials.

That it can be appreciable is shown by calculating response probabilities and latencies according to the general formulae:

$$P = \sum_i a_i P_i, \qquad m^{(r)} = \sum_i a_i P_i m_i^{(r)} \Big/ \sum_i a_i P_i,$$

where a_i is a frequency value (or density function in the continuous case) of the distribution (heterogeneity) of p such that $\Sigma a_i = 1$ and P_i and $m_i^{(r)}$ are the associated response probabilities and rth moments. For example, one result for heterogeneity in the form of a binomial distribution of parameter p with a mean value of 0·6 and $k=4$ in the random walk process gives correct and incorrect latencies of 9·5 and 10·7 respectively for a correct response probability of 0·92 in contrast to the equal latency value of 13·4 in the usual case. The lower overall latency is due, apparently, to the imbalance in response frequency and latency resulting from heterogeneity and this is one aspect of a more general influence upon the shape of the LPF. To investigate this problem, it is necessary to consider simplified examples of LPFs and it would seem that symmetrical or near-symmetrical LPFs will be skewed to the left; this is expected from the results for p-bias described above, which is a special case of heterogeneity with the distribution taking two values according to stimulus position.

An important point is that the effect of heterogeneity is to considerably increase the skew and kurtosis of latency distributions. This would appear due to the fact that the components of these heterogeneous distributions have a common minimum latency (i.e. threshold value) in the discrete time sense. Latency distributions from psychophysical discrimination tasks do not appear to have very high values of these statistics, in contrast to simple or choice reaction time distributions, and this feature should be taken into account: it will not be done so here, but elsewhere some empirical comparisons will be made.

In contrast, heterogeneity of the threshold value, k, results in a tendency for incorrect times to be shorter than correct. This arises because the longer responses, which are associated with the higher values of k, will have a greater probability of being correct. This particular form of heterogeneity could have an important application to the interpretation of differences between latencies in accuracy-instructed and speed-instructed tasks. For, if the result of a speed instruction is to introduce a heterogeneous and lowered threshold value into the discrimination process it becomes possible to account for the reversal of latency difference (dominant response having a longer latency) in some speed orientated subjects—e.g. see discussion of Greenberg's results in Coombs (1964, p. 535) and Wollen's (1963) fast responders discussed below. This is illustrated in Fig. 7 for $k=1$ and 2 in the accumulator, where all correct responses are now longer than incorrect. This result will be considered further below.

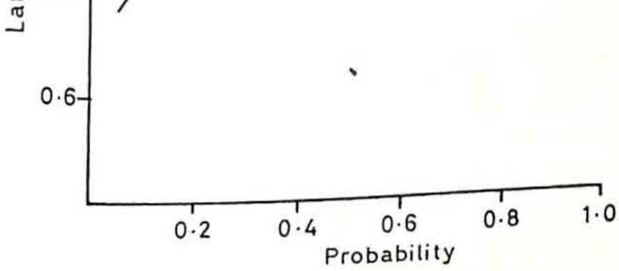


FIGURE 7. Theoretical LPF for an accumulator process with $k=1$ or 2 occurring with equal probability.

It should be noted that the k-bias discussed earlier is not an example of this form of heterogeneity which implies an equal value of k at any particular trial, but with this fluctuating between trials. In fact, this fluctuation may be related to variations in 'attention' and this possibility will also be discussed later.

4.3. *Decrease in Threshold Value*

If the threshold for a response is assumed to decrease with time within each trial, it is now by no means obvious that the effect is to produce shorter correct latencies, since a high threshold is associated with a longer latency but greater probability of being correct. It will suffice to illustrate that it does so in the special case of the random walk process, in which the value of k is assumed to decrease by 1 with each time interval that the process exceeds $k-1$ intervals without a response being achieved. For consider the probabilities of a correct response at times $k, k+1, \ldots$:

k: p^k (k steps in one direction.)

$k+1$: kp^kq (The process has taken one step, in one of k possible ways, in the incorrect direction with the result that the threshold has decreased by 1, with k steps in the correct direction still being necessary.)

$k+2$: $\binom{k+1}{2} p^kq^2$ (Two 'backward' steps with decrease of threshold to $k-2$.)

And so on until

$2k-1$: $\binom{2k-2}{k-1} p^kq^{k-1}$, when $k-1$ correct and incorrect pre-decisions have resulted in the next pre-decision being necessarily decisive, the threshold having been reduced to 1. No latency will occur which is longer than this value.

In the above case the probabilities resulting are those of the accumulator process with parameters p and k (e.g. see Audley & Pike, 1965) and they are, of course, similar for the incorrect response. Hence the process becomes identical to an accumulator and will possess the latency difference feature of the latter. Simulation of other forms of decrease with both the random walk and runs systems has also resulted in a latency difference, or in its increase when it is already present in a basic model.

One reason for describing a choice process in terms of a decreasing threshold is to take into account the possibility of subjects attempting to trade time with accuracy. It appears plausible that at least some subjects will be willing to do this and it is a conclusion of the RT studies of Snodgrass *et al.* (1967) that such a trading occurs and should be further investigated.

5. SOME EXPERIMENTAL RESULTS

Considering only the mean latencies, the results of previous work in psychophysical discrimination, such as that of Henmon (1911) (in an experiment with a single level of difficulty), Kellogg (1931), Laberge (1962) and Wollen (1963), appear to be in favour of the stochastic process theory of choice behaviour insofar as the correct latencies are generally shorter than incorrect. Results from some related studies are similar, especially those in which the subjects have to make a response along some hypothetical 'gradient' of generalization. For example, this situation has been studied for a series of vibratory stimuli by Gibson (1939), who found that the average latency of wrong responses was shorter and assumed this was the case, contrary to expectation, because long latency false responses had been inhibited. Later, Mednick (1958), in a similar situation, confirmed Gibson's prediction that if errors were not emphasized the expected negative correlation (i.e. between latency and probability) will be found.

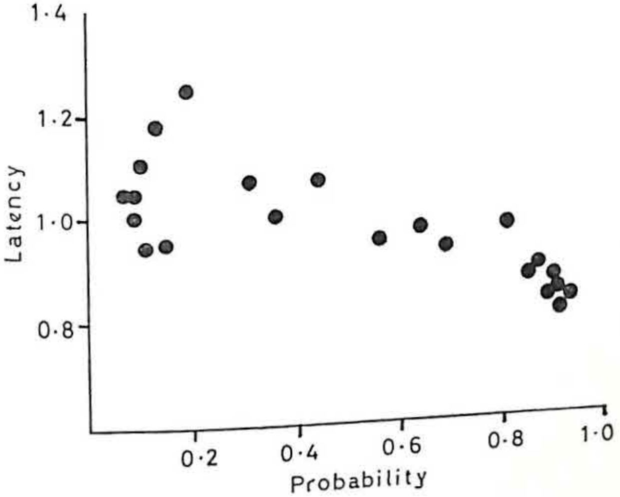
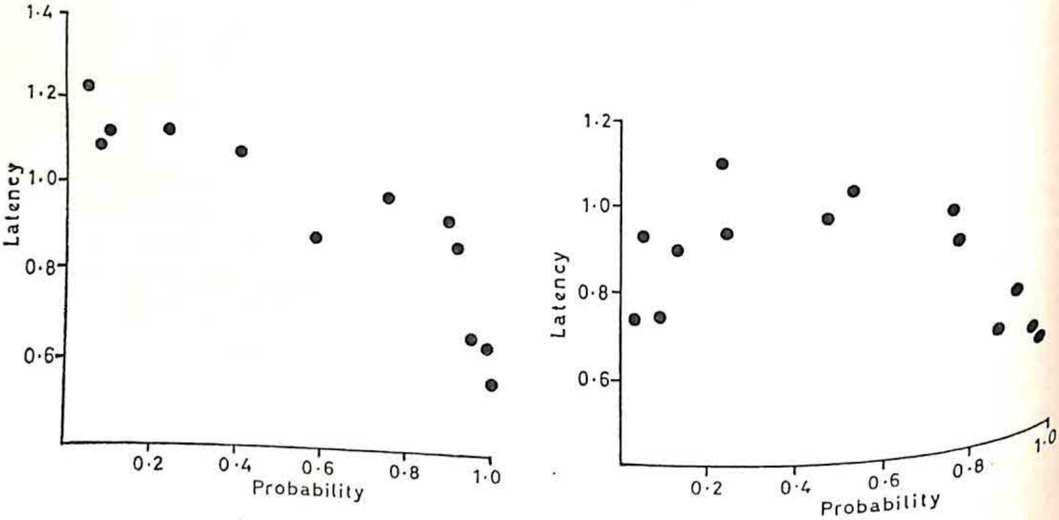


FIGURE 8. LP graph after Cross & Lane (1962). Points derived from graphs giving probability of response and latencies as functions of stimulus intensity for 20 subjects—two response modes combined in a single stimulus auditory discrimination task.

Buss (1950) had previously found the inverse relation in a study of concept formation and generalization. Pierrel & Murray (1963) found that each of eight subjects had shorter correct latencies in a weight-lifting task. Cross & Lane (1962) found similar results for 20 subjects and an LP graph may be derived from their data as presented in Fig. 8. The large variability of latencies for low probabilities is well demonstrated in these results.

The general theory of signal detectability has not specified how response latencies may be incorporated within its applications to human behaviour. This is unfortunate, since there clearly exists a correspondence between psychophysical discrimination and the signal-detection situation (cf. Treisman & Watts, 1966). The choice model of Laming (1962) is partially based on the classical signal-detection theory and it should not be difficult to construct latency models in which the theory plays an intrinsic part (i.e. with the response probabilities given directly by that theory). A very simple version is proposed by Audley & Mercer (1968) and is sufficient to indicate that if response latency varies inversely with the ' distance ' of the decision point from the detection criterion, then a continuously decreasing LPF is generated for the S + N and N latency distributions as these vary in distance from the criterion. Some latency results have been obtained in the detection situation: Sekuler (1965, 1966), in a study in which the signal was masked by an immediately following noise stimulus, obtained shorter latencies for correct ' Yes ' responses but no difference (at least in two of three subjects) in latency for ' No ' responses; Carterette *et al.* (1965) obtained, in the equiprobable case, shorter correct latencies for Signal and shorter incorrect latencies for Noise for two subjects, but the opposite results for a third subject. It would thus seem that bias effects are strong in this situation. Wolfendale (1957) has found that latency distributions in signal detection are not exponential in form but fit better a decision system based 'on runs of classified information from discrete observations '.

With regard to the LPF it will be observed that an important point of argument is whether or not the first half of it, for $P<0{\cdot}5$, falls below the value for $P=0{\cdot}5$. That it may do so in some cases is clearly shown in the data of Wollen (1963), who also made use of the effects of generalization in an experiment involving discrimination of spatial extent. In general, for experimentally



(*a*) (*b*)

FIGURE 9. LP graph derived from data given by Wollen (1963). (*a*) Mean of median latencies for 10 slow responders in unbiased group. (*b*) Mean of median latencies for 10 fast responders in unbiased group.

unbiased groups, latency increased as frequency decreased, but it was apparent to Wollen that the results could usefully be separated into 'slow' and 'fast' responders. The biased groups yielded data which give LP curves intermediate in shape between the latter two response groups. It will be seen immediately that the slow responders have average latencies which conform well with the form of the LP graph of the accumulator process, whereas the fast responders have a form indicative of a low *k* (or biased *k*) runs process.

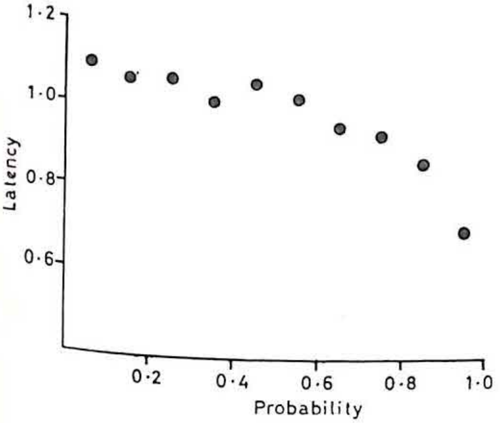


Fig. 10

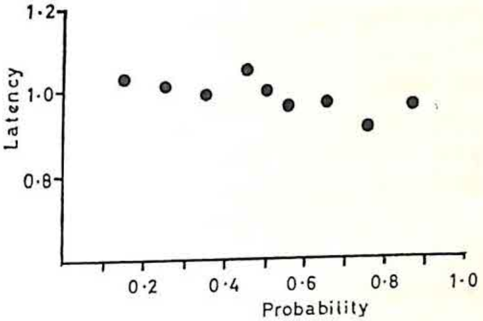


Fig. 11

FIGURE 10. LP graph for five subjects in length discrimination. Points are averages of mean latencies within intervals of 0·1 along *P* axis. Each point based on several cases for each subject who each received many trials with differing difficulty of discrimination.

FIGURE 11. LP graph for 15 subjects in a numerousness discrimination. Points give averages as in Fig. 10, and each interval contains several cases based on a total of approximately 112 trials for each subject with four conditions of difficulty. From Pike (1968).

Pickett (1967), in a texture discrimination task, found that response latency has a general inverse relation (64 out of 88 cases) with response probability and presents graphs similar to those presented here. He appears to be in favour of 'optional stopping' models for discrimination, and these would be formally similar to the stochastic models. Given in Figs. 10 and 11 are some results obtained by the writer which will be fully reported elsewhere. Fig. 10 gives the average LP graph for five subjects in a task involving discrimination of spatial extent, the details of which are given in Pike (1968). For this series of experiments 41 out of 54 pairs of correct/incorrect responses had shorter correct latencies and if the results for stimuli of equal difficulty (usually right/left pairs) are averaged then only one out of 27 has correct latencies longer. Fig. 11 gives the graph for 15 subjects in a numerousness discrimination (randomly spaced dots) when only 18 of 29 pairs had shorter correct latencies, as indicated by the flat LP graph.

6. DISCUSSION

The form of the empirical graphs presented here would not appear to fit those models of discrimination which would postulate some kind of ' measurement ' process insofar as those predict longer correct latencies; the evidence given by the difference between the mean times correct and incorrect or the LP graphs weighs heavily against such an interpretation. It is, however, possible to see how the observed time difference could be explained by modifying this system in ways similar to those discussed above for stochastic systems. Nevertheless, the form of these modifications will depend upon the precise description of such a system and it will, in general, be difficult to bring them about without introducing probabilistic parameters which will transform these models into stochastic ones in many ways similar to those discussed here.

It is also very doubtful if much credence can be given to the ' oscillating response tendency ' type of model which may be derived from the Hull–Spence system. As discussed in the introduction, the fact that there is no ' indecision ' in this model (in the competitive sense) results in geometric distributions of latencies with equal response times and there is little appearance of symmetry in the LP data illustrated here. The point may therefore be made that any model which attempts to describe behaviour at choice-points should represent the underlying indecision. This model is, of course, already stochastic in form and modifications would only make it resemble more the competing subsystem type discussed here.

It is clear that the stochastic models representing indecision give a reasonable first approximation to latency data. It is also fairly obvious from LP data that the random walk model in its simple form cannot be accepted because of its symmetrical LP curve. An interesting variation of this model has been proposed by Sekuler (1965), based on McGill's (1963) counter model for reaction times, in order to obviate the difficulties present in the (generalized) random walk of Stone (1960). He attempts to explain the data features mentioned previously, but the modifications necessary to this model to explain the fact that correct and incorrect ' No ' responses have equal latencies does not appear to have been satisfactory since he has not developed this particular approach (see Sekuler, 1966). He is clearly aware of the problem of bias present in these situations and advocates the further study of it. The study of Carterette *et al.* (1965) has further high-lighted this problem with the data of two of their three subjects being clearly biased in opposite directions to the third. These workers appear to account for their results in terms of a two-dimensional random walk (apparently aware that the one-dimensional latency model will not hold), but it is not at all clear how the model combines in a satisfactory manner with the detection theory.

Carterette *et al.* (1965) also criticize the Stone (1960) and Laberge (1962) models in terms of frequency distribution type; their latency data are generally of Pearson distribution Types I and VI (beta) or Type IV and they claim that the random walk and accumulator distributions are Type III (gamma). A very similar criticism is made of the use of the gamma distribution in general by

Snodgrass *et al.* (1967), who reckon that the Laberge (accumulator) model is a 'discrete analog of the gamma, the negative binomial'. But the latter authors also admit what is a very important point as far as any criticism based on the gamma distribution is concerned: that this Pearson type corresponds to an exact relation among the moments and there is no way of determining how close an empirical distribution approaches it. This is related to the fact that the empirical values of skew and kurtosis indices, which form the critical values of the Pearson distribution types, are generally biased as estimates of their theoretical values and the classification is sensitive to small changes in these values. (The approach advocated by Taylor (1965) in terms of cumulants may be found more useful in this respect.) It is also the case that for the lower values of k in the stochastic models the latency distributions are only approximately gamma and in the discrete-time accumulator are well within the beta classification. In view of these and other uncertainties involved in the fitting of empirical distributions by means of the Pearson classification (cf. Kendall & Stuart, 1958), it will not be discussed here in relation to the data obtained by the author except to mention that averaged skew and kurtosis indices generally indicate beta distributions (see Pike, 1967). It may be suggested that comparisons with computer simulation of theoretical models may, in future, yield better tests of those theoretical models than direct comparisons with their theoretical features, since simulated 'data' yield the expected statistics for given 'trial' numbers.

It would appear from the empirical LP data presented here that the most suitable model for psychophysical discrimination is the accumulator, particularly if only simple models (i.e. unbiased, homogeneous and fixed threshold) are considered, the form of the LP curves indicating this process rather than any other. The writer does not intend to conduct an exercise in curve-fitting at the present because the scope for modifications in the theoretical processes is so great: apart from those modifications discussed above, there are more empirically based ones such as the introduction of initial (i.e. non-response) time intervals or the consideration of aberrant fast or slow responses, although the former is probably not so important in discrimination times under accuracy instructions as in reaction time studies.

However, it is clear that the results do not unambiguously support the accumulator model. Some of the individual data of Pike (1968), Laberge (1962), the fast responders of Wollen (1963) and the fact that shorter incorrect latencies occur in speed-orientated subjects (cf. Coombs, 1964) indicate that the low probability end of the LPF may sometimes go below the value for $P = \frac{1}{2}$. Laberge suggests that the explanation lies in an artifactual fast-response process, Coombs (1964, p. 535) suggests a 'speed vs. accuracy set' and Wollen merely suggests the combination of slow and fast responders is artifactual. These results may be explained in the context of the accumulator model, as indicated previously, by supposing that 'fast' responders are using lowered and varying thresholds (heterogeneity in k). There are therefore two possible interpretations of the trading between accuracy and time in response latencies: one is that subjects may lower response ‘ thresholds ’ during a trial; the other that subjects may have ‘ thresholds ’ varying between trials, with the extent of this trading differing between subjects. The latter (heterogeneity) explanation is the more satisfactory since it predicts the speed vs. accuracy results, whereas the decreasing threshold version does not. (In the random walk, in fact, it predicts that slower responders, with little or no decrease, would have the more symmetrical LP curves.) Other explanations may be given in terms of observers operating with an accumulator process but using a ‘ runs ’ criterion if discrimination becomes protrated during a trial. This may have an interesting physiological interpretation which will be discussed elsewhere.

Finally, there is one particular explanation of the difference between correct and incorrect latencies in difficult discriminations which demands investigation because of its generality. This concerns the effects of ‘ attention ’ or ‘ concentration ’ in this situation. Variation of this within a set of trials may quite easily account for shorter correct latencies if it is assumed that the trial during which the subject is attending better is more likely to be correct and faster. This is an empirical fact if the assumption is correct and need not relate to theoretical heterogeneity of the type discussed above. In fact, there are now indications from physiological work (e.g. Fuster, 1958; Davis, 1964; Haider, 1967; Wilkinson, 1967) that something of this kind may occur; but whether it can account for all of the LP data is another matter.

ACKNOWLEDGEMENTS

The research reported in this paper was, in part, sponsored by the United States Air Force Office of Scientific Research, through the European Office of Aerospace Research, United States Air Force, when the author was at University College London. The author would like to thank Professor R. J. Audley for very useful criticisms and comments on an earlier draft of this paper.

REFERENCES

AUDLEY, R. J. (1960). A stochastic model for individual choice behaviour. *Psychol. Rev.* **67**, 1–15.

AUDLEY, R. J. & MERCER, A. (1968). The relation between decision time and relative response frequency in a blue–green discrimination. *Br. J. math. statist. Psychol.* **21**, 183–192.

AUDLEY, R. J. & PIKE, A. R. (1965). Some alternative stochastic models of choice. *Br. J. math. statist. Psychol.* **18**, 207–225.

BOWER, G. H. (1959). Choice-point behaviour. In R. R. Bush & W. K. Estes (eds.), *Studies in Mathematical Learning Theory*, chap. 6. Stanford: Stanford University Press.

BUSS, A. H. (1950). A study of concept formation as a function of reinforcement and stimulus generalization. *J. exp. Psychol.* **40**, 494–503.

CARTERETTE, E. C., FRIEDMAN, M. P. & COSMIDES, R. (1965). Reaction-time distributions in the detection of weak signals in noise. *J. acoust. Soc. Am.* **38**, 531–542.

CARTWRIGHT, D. & FESTINGER, L. (1943). A quantitative theory of decision. *Psychol. Rev.* **50**, 595–621.

Coombs, C. H. (1964). *A Theory of Data*. New York: Wiley.

Cross, D. V. & Lane, H. L. (1962). On the discriminative control of concurrent responses. *J. exp. Analysis Behav.* **5**, 487–496.

Davis, H. (1964). Enhancement of evoked cortical potentials in humans related to a task requiring a decision. *Science, N.Y.* **145**, 182–183.

Estes, W. K. (1950). Towards a statistical theory of learning. *Psychol. Rev.* **57**, 94–107.

Estes, W. K. (1960). A random walk model for choice behaviour. In K. J. Arrow, S. Karlin & P. Suppes (eds.), *Mathematical Methods in the Social Sciences*, pp. 265–276. Stanford: Stanford University Press.

Feller, W. (1957). *An Introduction to Probability Theory and its Applications*, vol. 1. New York: Wiley.

Fuster, J. M. (1958). Effects of stimulation of brain stem on tachistoscopic perception. *Science, N.Y.* **127**, 1150.

Gibson, E. J. (1939). Sensory generalization with voluntary reactions. *J. exp. Psychol.* **24**, 237–253.

Haider, M. (1967). Vigilance, attention, expectation and cortical evoked potentials. In A. F. Sanders (ed.), *Attention and Performance*. Amsterdam: North-Holland.

Henmon, V. A. C. (1911). The relation of the time of a judgement to its accuracy. *Psychol. Rev.* **18**, 186–201.

Hull, C. L. (1943). *Principles of Behaviour*. New York: Appleton-Century-Crofts.

Kellogg, W. M. (1931). The time of judgment in psychometric measures. *Am. J. Psychol.* **43**, 65–86.

Kendall, M. G. & Stuart, A. (1958). *The Advanced Theory of Statistics*, vol. 1. London: Griffin.

LaBerge, D. (1962). A recruitment theory of simple behaviour. *Psychometrika* **27**, 375–396.

Laming, D. R. J. (1962). A statistical test of a prediction from information theory in a card-sorting situation. *Q. J. exp. Psychol.* **14**, 38–48.

Laming, D. R. J. (1963). Human choice-reaction times. (Unpublished Ph.D. thesis, University of Cambridge.)

Laming, D. R. J. (1968). *The Information Theory of Choice-Reaction Times*. London: Academic Press.

Luce, R. D. (1960). Response latencies and probabilities. In K. J. Arrow, S. Karlin & P. Suppes (eds.), *Mathematical Methods in the Social Sciences*. Stanford: Stanford University Press.

McGill, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. 1. New York: Wiley.

Mednick, S. A. (1958). Gradients of latency in a generalized voluntary response. *Am. J. Psychol.* **71**, 752–755.

Pickett, R. M. (1967). Response latency in a pattern perception situation. *Acta psychol.* **27**, 160–169.

Pierrel, R. & Murray, C. S. (1963). Some relationships between comparative judgement confidence and decision-time in weight lifting. *Am. J. Psychol.* **76**, 28–38.

Pike, A. R. (1966). Stochastic models of choice behaviour: Response probabilities and latencies of finite Markov chain systems. *Br. J. math. statist. Psychol.* **19**, 15–32.

Pike, A. R. (1967). Stochastic interpretations of basic choice behaviour. (Unpublished Ph.D. thesis, University of London.)

Pike, A. R. (1968). Latency and response probability in simple discrimination. (In preparation.)

Rabbitt, P. M. A. (1966). Errors and error correction in choice response tasks. *J. exp. Psychol.* **71**, 264–272.

Rabbitt, P. M. A. & Phillips, S. (1967). Error-detection and correct latencies as a function of S-R compatibility. *Q. J. exp. Psychol.* **19**, 37–42.

Schouten, J. F. & Bekker, J. A. M. (1967). Reaction time and accuracy. In A. F. Sanders (ed.), *Attention and Performance*. Amsterdam: North-Holland.

Sekuler, R. W. (1965). Signal detection, choice response times and visual backward masking. *Can. J. Psychol.* **19,** 118–132.

Sekuler, R. W. (1966). Choice times and detection with visual backward masking. *Can. J. Psychol.* **20,** 34–42.

Snell, J. L. (1965). Stochastic processes. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. 3. New York: Wiley.

Snodgrass, J. G., Luce, R. D. & Galanter, E. (1967). Some experiments on simple and choice reaction time. *J. exp. Psychol.* **75,** 1–17.

Spence, K. W. (1954). The relation of response latency and speed to the intervening variables and N in S-R theory. *Psychol. Rev.* **61,** 209–216.

Spence, K. W. (1960). *Behaviour Theory and Conditioning*. Englewood Cliffs, N.J.: Prentice-Hall.

Stone, M. (1960). Models for choice reaction time. *Psychometrika* **25,** 251–260.

Takada, Y. (1960). Reaction time and information in the discrimination of length of lines. *Ex. Jap. Psychol.* **9,** 14–24.

Taylor, D. H. (1965). Latency models for reaction time distributions. *Psychometrika* **30,** 157–163.

Treisman, M. & Watts, F. R. (1966). Relation between signal detectability theory and the traditional procedures for measuring sensory thresholds. *Psychol. Bull.* **66,** 438–454.

Wiener, N. (1961). *Cybernetics*. New York: M.I.T.

Welford, A. T. (1960). The measurement of sensory-motor performance. *Ergonomics* **3,** 189–229.

Wilkinson, R. T. (1967). Evoked response and reaction time. In A. F. Sanders (ed.), *Attention and Performance*. Amsterdam: North-Holland.

Wolfendale, G. L. (1967). Decision times in signal detection. *Acta psychol.* **27,** 154–159.

Wollen, K. A. (1963). Relationship between choice time and frequency during discrimination training and generalization tests. *J. exp. Psychol.* **66,** 474–484.

Vol. 21
Part 2
183-192

The British Journal of Mathematical and
Statistical Psychology

November
1968

THE RELATION BETWEEN DECISION TIME AND RELATIVE RESPONSE FREQUENCY IN A BLUE–GREEN DISCRIMINATION

By R. J. Audley and Anne Mercer
University College London

Using a two-choice blue–green discrimination, decision times were measured for stimuli yielding relative response frequencies of approximately 0·10, 0·50, 0·90 and values close to zero and unity. The results were shown to support the prediction of a so-called accumulator, or recruitment, model of choice. It is also shown that the results could be well accounted for by assuming that the decisions were based on fixed samples of information which vary in quality.

1. Introduction

Audley & Pike (1965) subsumed some models of choice in a simple general representation in order to facilitate comparisons among them. They hypothesized that in any choice situation, discrete units of information arriving randomly in time were stored according to the alternative they favoured. It was further assumed that an overt choice occurred when the contents of the stores satisfied certain specified conditions. Three models were given particular attention: (1) *the accumulator*—as soon as one store receives k units the appropriate response is made (Laberge's (1962) recruitment model); (2) *accumulator with inhibition*—the same decision rule applies, but the arrival of a unit of excitation in one store clears all other stores (Audley's (1960) 'runs' model); (3) *random walk*—a choice is made when the contents of a store exceeds that of all others by a fixed number of units k (a simplified analogue of the optimal sequential statistical scheme considered as a model for choice by Stone (1960) and Laming (1968)).

In sketching out a mathematical treatment of these models, Audley & Pike pointed out that the relation between the time to make a choice and the relative frequency with which that choice was made had a considerably different predicted form for the three models (shown in Fig. 1). In this respect, the particular experimental situation they envisaged was one in which the subject had to decide, on each of a series of trials, which of two stimuli was being presented to him, the stimuli appearing with equal frequency in a random order. The relative frequency of the alternative responses was to be manipulated by varying the difficulty of the discrimination. In this manner it was hoped to avoid biases in favour of one of the alternatives as happen, for example, when variations in the relative frequency of a response are achieved by making one stimulus more frequent. For one thing, the asymmetries in the theoretical decision processes which would usually be postulated to handle this latter situation (e.g. different k values for the different responses) considerably diminish the differences between the predictions from the alternative models. For another, bias

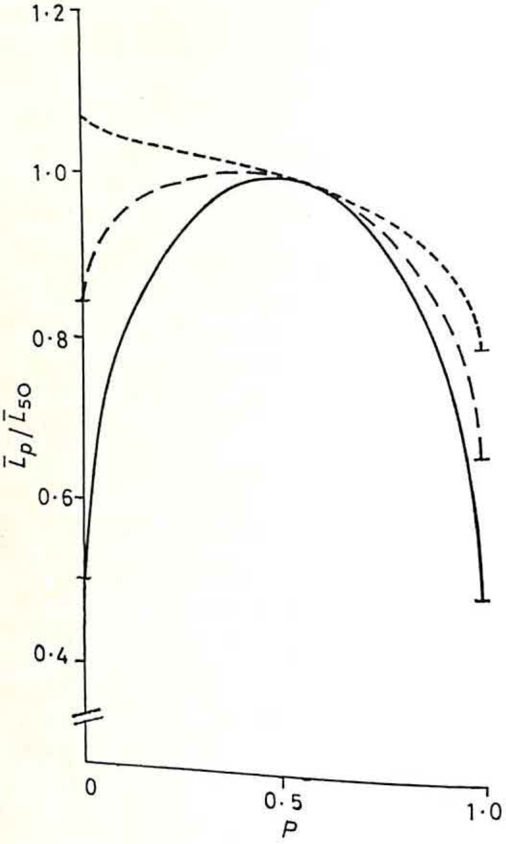


FIGURE 1. Theoretical prediction of relation between decision time and the relative frequency of a response ($k=2$). ——— = Random walk. - - - - = Accumulator; — — — — = Runs;

in favour of one response makes it likely that new factors associated with response sets may enter the experimental situation so as to weaken the relation between theory and experiment. Varying discriminability is not without its own shortcomings as a means for manipulating relative response frequency. In particular, for the particular kind of representation of choice discussed by Audley & Pike, it might be supposed that as discriminations become easier, not only is there a greater relative frequency of excitation in favour of one alternative (which these authors did take into account), but also the excitation might arrive at a higher rate. This last point needs further investigation. Here it will be sufficient to state that the results of the present experiment apparently do not require such an effect to be postulated.

The first author has been investigating this relation between time and relative frequency for individual subjects and these data will be reported subsequently. But it was thought desirable to obtain results from a number of individuals in such a way that they could be meaningfully averaged to throw light upon the general form of the relation in question. The experiment reported here was designed to obtain decision times for a few strategically selected relative frequencies of response. Those chosen were 0·10, 0·50 and 0·90, together with some observations on responses with relative frequencies close to zero and unity. By confining attention to these selected points it is possible to average subjects' times and still obtain theoretically interesting empirical results. Furthermore, by expressing an individual's times relative to that he takes for the response with a relative frequency of 0·50, it is possible to eliminate any effects due to differences in overall rate of responding.

The discrimination of colours was chosen as the setting for the experiment because it was hoped that the subject would bring to such a task already established and relatively stable perceptual templates for labelling the hue of a stimulus. This meant that there would be a greater chance of him being able to categorize each member of a whole range of different stimuli into two alternatives than could be achieved if he had been required to dichotomize some other dimension (say, intensity) which would have demanded considerable pretraining prior to the experiment proper.

The blue–green range was chosen because it is supposed to be relatively free of stimulus anchors (Lenneberg, 1957). It was therefore hoped that each trial would require a genuine categorization from the subject and would not merely involve the re-labelling of a number of relatively identifiable hues as being blue or green.

2. EXPERIMENT

Method The subjects were presented with hues in the blue–green range and required to decide whether these were blue or green and to indicate their choice by depressing one of two appropriately labelled morse keys. The time elapsing between the presentation of the stimulus and a key being depressed was the main dependent variable. Half of the subjects were also required to rate their confidence in each judgement. In order to ensure that this did not in itself affect the pattern of decision times, the remaining subjects were not required to make this additional rating.

Blue–green hues of seven different types were presented, the actual stimulus values being separately gauged for each subject.

Three of the mixtures were chosen so that a subject would produce respectively 10%, 50% and 90% blue responses. This was achieved by using a ' staircase ' procedure (e.g. see Wetherill & Levitt, 1964). In the 90% condition, for example, if the subject produced six consecutive blue responses to presentations of this type, then on the next occasion the stimulus mixture was set one step closer to green; if a green response was ever produced, a step was made towards blue. This procedure leads to a series of stimuli which will stabilize at a point where on 50% of the trials six blue responses are made and for the other 50% there are fewer than six. If P is the relative frequency of blue responses at this stabilization point, then $P^6 = 0{\cdot}5$ and therefore $P = 0{\cdot}89$.

Two of the mixtures were set farther out towards the extremes so that they would be expected to yield virtually 100% consistent decisions. The remaining two were put at even more extreme values. The seven conditions will be labelled **0**, 0, 10, 50, 90, 100, **100** respectively. The stimulus conditions were presented in random order to the subject in five blocks of 32 trials. Each block contained eight trials of conditions 10, 50 and 90 and two trials of each of the remaining conditions.

Subjects Ten undergraduate students of University College London.

Apparatus A Pickford–Nicholson anomaloscope (Pickford, 1949) fitted with a blue filter (wavelength 400 mμ) and a green (wavelength 520 mμ) was used as the source of a range of blue–green hues. A 12 in. diameter perspex wheel was mounted on the shaft of the hue-mixing control of the anomaloscope so that 78 steps could be obtained between the blue and green extremes of the colour range. The subject sat 6 ft. away from an aperture 0·7 in. in diameter cut in a brown hardboard screen. A warning light was situated 4·5 in. below the aperture. A shutter behind the aperture controlled the presentation of the stimulus which was terminated by the subject's response. The selection of a stimulus hue on each trial was carried out manually by an experimenter in the same room as the subject, but hidden from him. The presentation of the stimuli was controlled by a second experimenter in an adjoining room.

The subject was provided with two morse-type keys to signal his decision and these were placed under his right and left index fingers, respectively. A microphone was also provided to enable the experimenter in the adjoining room to note down the judgements of confidence required from some subjects. This experimenter also recorded the subject's choice and decision times on a Dekatron timer to the nearest 10 millisec.

Procedure As subjects varied greatly in their naming of any particular hue, an approximate 50% point for each subject was found by the method of limits with six settings of the stimulus. The starting values for the 10% and 90% conditions were made at three arbitrary scale units from this approximate mid-point. The 0, 100 conditions were placed three further units towards the extremes, and the **0, 100** were set at either four or eight units farther out. (For five subjects, their most extreme point was towards the green; for the others it was towards the blue.)

Stimuli were presented automatically at a rate of one every 14 sec. The warning signal came on briefly and was followed 1 sec. later by the stimulus.

Subjects were instructed that they had to decide whether the light presented on each trial was blue or green, and to respond as quickly as possible without being inaccurate. Five of the subjects were asked to rate their confidence in their judgements on a 5-point scale from 55 (very uncertain) to 95 (highly confident).

3. RESULTS

The data were analysed in two ways. In the first, all the data for a given condition were used to calculate the average decision time. In the second, the only data used were those for the stimulus value in the condition which produced a proportion of correct responses closest to the target relative frequency. There was hardly any difference in the relation between decision time and relative frequency obtained by the two methods, so the results presented here are those based on the first. There were no marked differences between the properties of blue and green responses and so the data for these were pooled. The average response probabilities associated with the conditions were 0·02, 0·07, 0·014, 0·50, 0·86, 0·93 and 0·98.

Each subject's time was standardized by expressing it as a ratio to his time in condition 0·50. These ratios were then averaged. In Fig. 2 they are plotted together with the theoretical curve for the accumulator model (with the threshold value $k=2$) which describes the data quite well, although the empirical decision time for correct probability is rather higher than that theoretically expected (the monotonic trend in the ratios is significant with $P<0{\cdot}001$ using Page's L test). For the accumulator model, the expected decision time ratios for the 10 and 90 conditions are: for $k=2$, 1·05 and 0·91; for $k=3$, 1·06 and 0·89; and for $k=10$, 1·05 and 0·90. The obtained values for each subject are shown in Table 1 and they have mean values of 1·16 and 0·88 respectively. These values are not significantly different from those predicted from the accumulator model with $k=2$ ($t=1{\cdot}79$, $P>0{\cdot}10$; and $t=0{\cdot}94$, $P>0{\cdot}25$, respectively). The ratio for condition 10 is, however, significantly larger than that predicted from the runs model with $k=2$ ($t=2{\cdot}90$, $P<0{\cdot}001$) and therefore *a fortiori* than the theoretical value for the random walk model. For condition 90, the observed ratio does not differ significantly from that expected on the runs model ($t=0{\cdot}37$, $P>0{\cdot}25$).

The analysis of the response time data for relative frequencies near 100% presents special problems which will be discussed more fully elsewhere in connection with the results of another experiment.

There is nothing of special interest to report about the confidence judgements except that their presence did not unduly influence the subject's decision time. For condition 10, both groups of subjects gave the same average decision time ratio; and for condition 90, the average ratios were 0·90 for those subjects requested to give confidence ratings and 0·86 for the remainder.

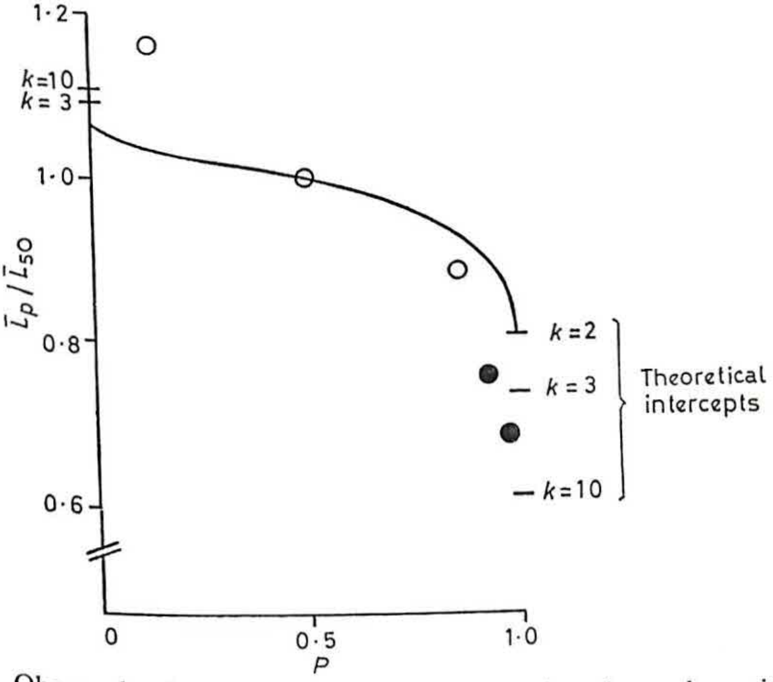


FIGURE 2. Observed values of decision time ratios plotted together with theoretical prediction from accumulator model.

TABLE 1. DECISION TIME RATIOS FOR INDIVIDUAL SUBJECTS

Subject	Observed response probability	Decision time ratio ($\bar{L}_p/\bar{L}_{50}$)
N.O.	0·125	0·99
	0·875	0·86
L.B.	0·190	1·19
	0·810	0·98
A.B.	0·125	1·19
	0·875	0·99
Z.K.	0·140	1·23
	0·860	0·75
C.G.	0·125	1·16
	0·875	0·93
D.P.	0·138	0·88
	0·865	0·74
G.M.	0·180	1·22
	0·825	0·97
M.R.	0·125	1·04
	0·875	0·98
A.P.	0·125	1·62
	0·875	0·64
L.P.	0·190	1·00
	0·810	0·85

4. DISCUSSION

The technique employed here has yielded clear evidence on the relation between decision time and relative response frequency. The times for the 10 and 90 conditions clearly have a ratio to that for the 50 condition which is closest to that predicted from the accumulator model (or recruitment model (Laberge, 1962)); and the ratios for the lower relative frequency are clearly incompatible with the other schemes discussed by Audley & Pike (1965). It is true that the 0 and 100 conditions occasionally yielded some fast times for relatively infrequent responses which are incompatible with the model. But these short latency errors in easy discriminations have recently been identified as a special class (Ollman, 1966; Yellott, 1967). Although they need to be studied further, they were not sufficiently frequent in the present experiment to suggest that we should modify our conclusions.

In calculating the observed decision time ratios, it was assumed that the total time that the subject took in making his choice was attributable entirely to an underlying decision process. In fact, there is likely to be a considerable contribution from other sources, such as peripheral conduction times. The effect of taking this into account would be to increase slightly the discrepancy between the data and the theoretical predictions from the accumulator model. However, these predictions have been based here on rather specific assumptions about the input to the decision-making mechanisms. The essence of a literal interpretation of the model is a race between two systems independently waiting to achieve a threshold level of excitation (see also Gleitman & Caplin, 1968). An appropriate change in the assumptions about the input could alter the predicted decision time ratios in the direction of those observed.

The support for the accumulator model now changes the direction of the enquiry. For not only is the model capable of several interpretations, there is a rather different approach altogether which yields a similar theoretical result. On the former point, for example, Audley & Pike (1965) pointed out that a random walk decision model (the optimal statistical decision-making procedure) with boundaries which shrink in time has the properties of the accumulator. Especially in the present situation which involved extremely ambiguous stimuli, one might expect that, as time passes, subjects would rapidly lower their criterion concerning the level of evidence to be treated as decisive. Indeed, if the sequential statistical model is to describe successfully the present kind of data, then such shrinkage of the boundaries is required, for otherwise subjects would be expected always to express the same degree of confidence in their judgements. This certainly does not happen; and a poorer accuracy in performance accompanies the lower confidence ratings—the very reason why one may use the signal detectability model to describe this kind of discrimination data.[1] But it is also possible that the form of the present results is due to a considerable degree of trial to trial variation in the quality of the information that the subject is processing.

Let it be supposed that on the presentation of a stimulus the subject receives a limited sample of information, the usefulness of which varies from trial to trial: on some trials excellent information is available and the subject is readily and quickly able to determine on which side of the blue–green mid-point the stimulus falls; on others, poor information will demand a lengthy measurement process which in the end leads to a wrong choice. If attention is limited to simple order statistics it is possible to use an extension of the signal detectability model (Green & Swets, 1966) to derive predictions for such a model.

Let P–Blue be a stimulus which on its presentation elicits from the subject a proportion P of blue responses. Suppose that the samples of information received when a hue is presented can be represented as varying in some single dimension of quality—i.e. akin to the likelihood ratio abscissa of signal detectability theory. Fig. 3 shows how certain order statistics can be derived.

This figure shows the distribution of samples expected from three hues: a $P=0{\cdot}50$ mid-blue–green; and two other hues which produce response proportions symmetrically above and below this. The distributions are supposed to be normal with equal variance, here set equal to unity. It is also assumed that the mean of the $P=0{\cdot}50$ distribution is the position of the criterion above which a sample of information will always be called blue by the subject and below

---

[1] In this connection, it is of interest that the most highly motivated subject was an art student who had recently painted a study in blues and greens! Every decision was of great professional importance to him and one might expect him to maintain a high criterion over time; but the accumulator model still fitted his data well and there was no tendency towards the random walk form.

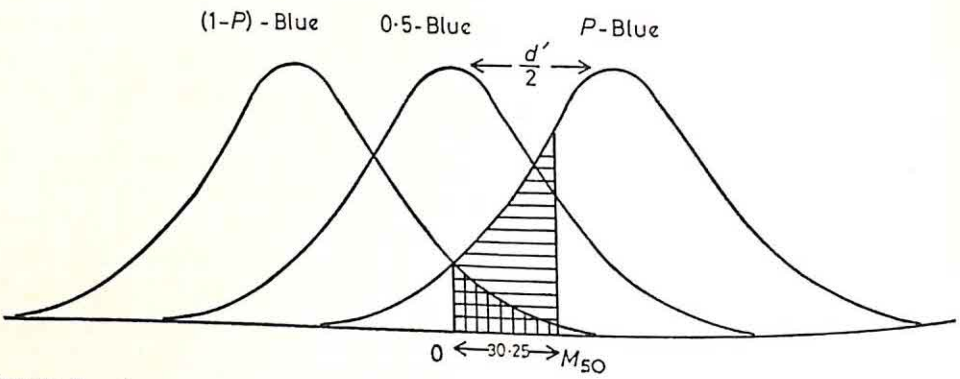


FIGURE 3. A representation of the hue discrimination task. Proportion of correct responses with decision time above $M_{50}=A/P$. Proportion of incorrect responses with decision time above $M_{50}=B/(1-P)$. Horizontal hatching $=A=$ Proportion of responses which are correct and longer than M_{50}. Vertical hatching $=B=$ Proportion of responses which are incorrect and longer than M_{50}.

which it will be called green. Consider the latencies of the blue responses made to the 0·5–Blue stimulus. Obviously, the median latency (M_{50}) accompanies these responses made to samples of information which are of median usefulness; and this occurs at a standard score from the mean of the distribution of $z_{0\cdot 25}=0\cdot 68$. Now consider responses to the P–Blue stimulus distribution. The area A of this distribution lying between the criterion and $z_{0\cdot 25}$ gives the proportion of responses which are both called blue and are based on poorer information than that requiring time M_{50} to process, and hence have decision times longer than M_{50}. This area is

$$A=P-\int_{0\cdot 68-d'/2}^{\infty}\exp(-z^2/2)dz,$$

where d' is the distance between the mean of the P– and $(1-P)$–Blue distributions, calculated from the expression:

$$P=\int_{-d'/2}^{\infty}\exp(-z^2/2)dz.$$

Therefore the proportion of blue responses to this stimulus which have latencies longer than M_{50} is A/P. Similarly the proportion of blue responses to the $(1-P)$–Blue stimulus which have decision times longer than M_{50} is $B/(1-P)$, where

$$B=(1-P)-\int_{0\cdot 68+d'/2}^{\infty}\exp(-z^2/2)dz.$$

The graph of these proportions as P varies is shown in Fig. 3, together with the data points for $P=0\cdot 9$ obtained from the present experiment. Some of the order statistics are based on small samples, but the results indicate that the variability hypothesis could account for the kind of data obtained in the present experiment. If one considers predictions within a given condition, e.g. the proportion of error times longer than the median correct time, the fit of the model is very much better.

The form of the theoretical curve in Fig. 3 is obviously dependent on the assumed shapes of the distributions of stimulus information. Dr E. A. C. Thomas has very kindly looked at the predictions following from the assumption of Laplace and uniform distributions. In the former case, the curve is horizontal for response probabilities lying between zero and 0·5, and it then decreases monotonically. For the rectangular case, there is a short horizontal segment for low response probabilities and then a monotonic decrease.

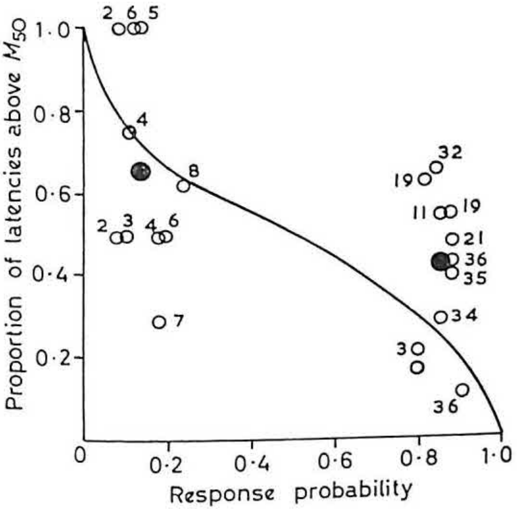


FIGURE 4. Latency predictions from statistical model. ● = Average.

A fuller discussion of the relative merits of the explanation of these data provided by the accumulator model and rival hypotheses requires further evidence which will be presented elsewhere. But the results of the present experiment and others which will be subsequently reported (see especially Pike, 1968) indicate that to a first approximation, the accumulator provides a good prediction of the relation between latency and probability in a two-choice discrimination task in which the alternative stimuli appear equally frequently.

ACKNOWLEDGEMENTS

This research was supported by a grant to the first author by the Medical Research Council. The authors are grateful for the helpful comments made about an earlier draft by Drs A. R. Pike, T. Shallice and E. A. C. Thomas.

REFERENCES

AUDLEY, R. J. (1960). A stochastic model for individual choice behaviour. *Psychol. Rev.* **67**, 1–15.

AUDLEY, R. J. & PIKE, A. R. (1965). Some alternative stochastic models of choice. *Br. J. math. statist. Psychol.* **18**, 207–225.

GLEITMAN, H. & CAPLIN, J. (1968). Choice probability as a function of the latency of the two individual response alternatives. (Personal communication.)

GREEN, D. M. & SWETS, J. A. (1966). *Signal Detection Theory and Psychophysics.* London: Wiley.

LABERGE, D. (1962). A recruitment theory of simple behaviour. *Psychometrika* **27,** 375–396.

LAMING, D. R. J. (1968). *The Information Theory of Choice-Reaction Times.* London: Academic Fress.

LENNEBERG, E. H. (1957). A probabilistic approach to language learning. *Behav. Sci.* **2,** 1–12.

OLLMAN, R. (1966). Fast guesses in choice reaction time. *Psychon. Sci.* **6,** 155–156.

PICKFORD, R. W. (1949). Individual differences in colour vision and their measurement. *J. Psychol.* **27,** 153–202.

PIKE, A. R. (1968). Latency and relative frequency of response in psychophysical discrimination. *Br. J. math. statist. Psychol.* **21,** 161–182.

STONE, M. (1960). Models for choice reaction time. *Psychometrika* **25,** 251–260.

WETHERILL, G. B. & LEVITT, H. (1965). Sequential estimation of points on a psychometric function. *Br. J. math. statist. Psychol.* **18,** 1–10.

YELLOTT, J. I. (1967). Correction for guessing in choice reaction time. *Psychon. Sci.* **8,** 321–322.

Vol. 21
Part 2
193–200

The British Journal of Mathematical and
Statistical Psychology

November
1968

DISCRIMINATIVE BINARY PREDICTION WITH REINFORCED CUE IDENTIFICATION

By Dominic W. Massaro
University of California, San Diego

and John W. Moore
University of Massachusetts, Amherst

Two tones, T_1 and T_2, differing in loudness, were employed as stimuli in a discriminative binary prediction task. Two responses, A_1 and A_2, predicting events E_1 and E_2 respectively, were available to the subjects. An A_1 response was correct with probability π_1 and π_2 on T_1 and T_2 trials, respectively. The present study assessed the effects of the covariation of π_2 and the similarity of T_1 and T_2 on $P(A_1|T_1)$, the probability of an A_1 response given a T_1 trial. Subjects were also required to identify the trial type (loud or soft) and were given partial feedback of identification responding. Partial feedback was sufficient to eliminate the effects of cue similarity on discriminative event prediction and also improved identification of trial types. The reinforcement effects indicated that subjects learn to behave appropriately to identified cues rather than learn to make specific motor responses.

1. Introduction

The experimental situation of discriminative binary prediction consists of a series of trials, each of which commences with the onset of one of two cue stimuli, T_1 or T_2. The probability of T_1 is β, and the probability of T_2 is $1-\beta$. Two responses, A_1 and A_2, predicting events E_1 and E_2 respectively, are available to the subject. On a T_1 trial, an A_1 response is correct with probability π_1, and an A_2 response is correct with probability $1-\pi_1$. On a T_2 trial, an A_1 response is correct with probability π_2, and an A_2 response is correct with probability $1-\pi_2$.

A previous experiment by Massaro *et al.* (1968) has shown that $P(A_1|T_1)$, the probability of an A_1 response given a T_1 trial, is dependent upon both π_2 and the similarity between T_1 and T_2. With values of $\pi_1=0{\cdot}8$ and $\pi_2=0{\cdot}8$, $0{\cdot}5$ or $0{\cdot}2$, the results indicated that $P(A_1|T_1)$ was: (1) a linear function of π_2 when the cues were highly similar, (2) independent of π_2 and equal to π_1 at an intermediate level of cue similarity, and a U-shaped function of π_2 when the cues were highly discriminable. In that study, in order to determine the similarity between T_1 and T_2, the subjects were required to indicate which of the two cues (loud or soft tones) was being presented before making their prediction of the event that would occur on that trial. However, the subjects were not given feedback regarding the accuracy of their identification of trial type. The present experiment is an assessment of the effects of variations of π_2 and cue similarity on $P(A_1|T_1)$ with partial feedback (25 or 75 per cent) about identification responses.

2. METHOD

Subjects The subjects were 144 University of Massachusetts undergraduates and they were assigned randomly to the experimental treatments.

Apparatus Up to four subjects were run at a time, each seated at a tabletop enclosure containing an Estes Straughan conditioning board, which consisted of a white centre warning light and two red event lights, each positioned above a spring-loaded lever switch. The centre of the board contained a 2 × 4·5 in. panel recessed 1·5 in. with two spring-loaded buttons mounted one above the other and labelled loud and soft. Tones were generated by a Hewlett–Packard Model 200 audio-oscillator and were presented over matched headphones with a continuous white masking noise. Experimental events were controlled by Lehigh Valley 1620 Probability Randomizers, Hunter Interval Timers, and relays. Events and responses were recorded on an Esterline–Angus event recorder.

Procedure The onset of a tone started a trial. The tone lasted 2·5 sec., during which the subjects were required to make a loud or soft identification response by pressing the respective button. At the end of the tone, the correct button was illuminated for 0·9 sec. on feedback trials. On both feedback and no feedback trials, the warning light was illuminated 0·9 sec. after tone offset. The light was left on for 1·5 sec. and the subjects made their prediction responses. The event light was illuminated for 0·9 sec. immediately following the end of the warning light. Hence each trial lasted 5·8 sec. The inter-trial interval was 4·2 sec. Each subject received 300 trials in which all stimuli, feedback and events were presented randomly by an appropriate setting on the probability randomizers.

Subjects were given the following instructions: 'You will be receiving two tones differing slightly in loudness over the headphones. Your first task will be to indicate which of the tones you are listening to. You will do this by pressing one of the two buttons. You will press the top button for the louder tone and the bottom button for the softer tone. Notice the buttons are labelled loud and soft. You are expected to guess if you are not sure of your decision. You will have 2½ sec. to make your choice. After 2½ sec., the tone will go off and one of the two buttons *may* come on, indicating whether the loud or soft tone did come on. Immediately after this, the white light on top of your panel will go on. Your second task is to *predict* which of the two red lights at the bottom of your panel will come on. As soon as possible after the white light comes on, you are to flip one of the two switches. After you have flipped a switch, one of the two red lights will show. If the red light above the switch you pressed comes on, you were correct. If the other one comes on, you were incorrect on that trial. Remember, you will always be told whether or not you have been correct on the second task. You will not always be given information regarding your indication of a loud or soft tone.'

Design Three levels of cue similarity (ΔI) were crossed with three stimulus contingent reinforcement schedules and two feedback conditions. The two feedback conditions were 25 and 75 per cent feedback. The intensity pairings of the two 800 Hz tones were 73–74·5, 73–76 and 73–79 db SPL, giving a 1·5, 3 and 6 db intensity differential (ΔI), respectively, for three decreasing levels of cue similarity. The three reinforcement schedules (π) were: (*a*) $\pi_1 = P(E_1|T_1) = 0{\cdot}8$ for all groups, and (*b*) $\pi_2 = P(E_1|T_2) = 0{\cdot}8$, 0·5 and 0·2 respectively for group 0·8–0·8, group 0·8–0·5 and group 0·8–0·2. Thus, there were eight subjects in each cell of a 3 × 3 × 2 between-groups factorial design for a total of 144 subjects. The analysis of variance of $P(A_1)$'s included the three between variables of ΔI, π and feedback, and the three within variables of cue (T_1 vs. T_2), feedback trial (yes or no), and trial block (three blocks of 100 trials).

3. RESULTS

Marginal statistics Fig. 1 shows that discrimination performance, $P(A_1|T_1) - P(A_1|T_2)$ ($F=180 \cdot 21$; d.f. $=1, 126$; $P<0 \cdot 001$), increased as the difference between π_1 and π_2 increased ($F=81 \cdot 48$; d.f. $=2, 126$; $P<0 \cdot 001$). Discrimination performance also increased over training ($F=19 \cdot 56$; d.f. $=2, 252$) and increased faster over training for larger π_1 and π_2 differences ($F=7 \cdot 00$; d.f. $=4, 252$; $P<0 \cdot 001$). The cue × ΔI interaction and their 15 higher order interactions were not statistically significant. This finding that cue similarity did *not* disrupt discrimination performance in any significant way indicates that partial feedback was sufficient to eliminate the effects of cue similarity found in previous studies (Massaro *et al.*, 1968).

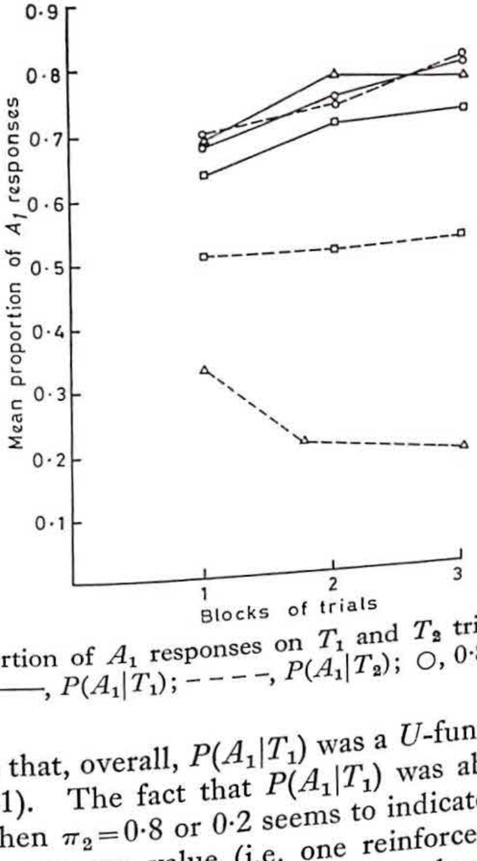


FIGURE 1. Mean proportion of A_1 responses on T_1 and T_2 trials as a function of trial blocks and π_2. ——, $P(A_1|T_1)$; - - - -, $P(A_1|T_2)$; ○, 0·8–0·8 group; □, 0·8–0·5; △, 0·8–0·2.

Fig. 1 also shows that, overall, $P(A_1|T_1)$ was a U-function of π_2 ($F=172 \cdot 7$; d.f. $=1, 126$; $P<0 \cdot 001$). The fact that $P(A_1|T_1)$ was about 5 per cent lower when $\pi_2=0 \cdot 5$ than when $\pi_2=0 \cdot 8$ or $0 \cdot 2$ seems to indicate that a discriminative stimulus with little or no cue value (i.e. one reinforced on a 0·5–0·5 basis) depressed $P(A_1|T_1)$ relative to a stimulus with cue value. This finding agrees with studies where the trial types were highly discriminable (Popper & Atkinson, 1958), except that the overshooting reported in that study was not found in the present experiment. This overall depression of $P(A_1|T_1)$ could be due to the randomness of the occurrence of feedback in the present study. For example, the subjects may have hypothesized that event occurrence was also dependent on whether feedback was given on a trial. This hypothesized contingency

TABLE 1. OBSERVED AND PREDICTED VALUES OF $P(A_{1,n+1}|T_{i,n+1}T_{j,n}A_{k,n}E_{l,n})$ FOR GROUP 0·8–0·2

('No feedback' trials; trial type correctly identified for the 25 per cent feedback condition pooled over levels of ΔI and trial blocks.)

				Frequency		Proportion	
$T_{i,n+1}$	$T_{j,n}$	$A_{k,n}$	$E_{l,n}$	Observed	Predicted	Observed	Predicted
1	1	1	1	365	380·9	0·867	0·905
1	1	1	2	84	89·0	0·737	0·746
1	1	2	1	59	52·1	0·634	0·560
1	1	2	2	17	15·4	0·436	0·395
1	2	1	1	10	9·9	0·400	0·395
1	2	1	2	63	61·0	0·578	0·560
1	2	2	1	93	91·7	0·756	0·746
1	2	2	2	430	433·4	0·898	0·905
2	1	1	1	42	45·1	0·089	0·095
2	1	1	2	20	34·8	0·146	0·254
2	1	2	1	50	57·2	0·385	0·440
2	1	2	2	15	20·0	0·455	0·605
2	2	1	1	14	13·3	0·636	0·605
2	2	1	2	44	47·1	0·411	0·440
2	2	2	1	30	28·5	0·268	0·254
2	2	2	2	50	45·6	0·104	0·095

$c=0{\cdot}314$; $c_2=0{\cdot}334$; $\chi^2=11{\cdot}20$.

TABLE 2. .OBSERVED AND PREDICTED VALUES OF $P(A_{1,n+1}|T_{i,n+1}T_{j,n}A_{k,n}E_{l,n})$ FOR GROUP 0·8–0·2

(Feedback trials; trial type correctly identified for the 75 per cent feedback condition pooled over levels of ΔI and trial blocks.)

				Frequency		Proportion	
$T_{i,n+1}$	$T_{j,n}$	$A_{k,n}$	$E_{l,n}$	Observed	Predicted	Observed	Predicted
1	1	1	1	623	599·7	0·941	0·906
1	1	1	2	107	104·6	0·728	0·711
1	1	2	1	74	63·9	0·692	0·597
1	1	2	2	12	8·7	0·546	0·394
1	2	1	1	19	10·6	0·704	0·394
1	2	1	2	4	59·7	0·040	0·597
1	2	2	1	93	81·1	0·816	0·711
1	2	2	2	415	412·2	0·912	0·906
2	1	1	1	57	54·8	0·098	0·094
2	1	1	2	28	39·0	0·207	0·289
2	1	2	1	25	38·3	0·263	0·403
2	1	2	2	17	20·0	0·515	0·606
2	2	1	1	21	21·8	0·583	0·606
2	2	1	2	41	47·5	0·348	0·403
2	2	2	1	46	35·2	0·377	0·289
2	2	2	2	48	50·6	0·089	0·094

$c=0{\cdot}383$; $c_2=0{\cdot}413$; $\chi^2=76{\cdot}85$.

The observed rank-orderings in Table 1 agree with the predictions of the model based on the subject making an *appropriate* response. Therefore, when trial types n and $n+1$ are identified as the same,

$$P(A_1|A_1E_1) > P(A_1|A_1E_2) > P(A_1|A_2E_1) > P(A_1|A_2E_2),$$

whereas when the trial types are identified as different, the rank-ordering is exactly opposite. Table 2 shows that reinforcement of identification responses does not disrupt this mode of response, since the rank-orderings are the same as for the trials without feedback. These results support the thesis that, in discrimination learning, subjects are reinforced for behaving appropriately to identified cues rather than for making specific motor responses.

4. CONCLUDING REMARKS

In discrimination learning, the reinforcement schedule in operation for one trial type does affect responding to the other trial type. A stimulus with some cue validity enhances appropriate responding to the other stimulus compared to a stimulus with no cue value (i.e. one reinforced on a 0·5–0·5 basis). When the subject is reinforced for behaving appropriately (i.e. predicting the most frequent event given the trial type) he will be more likely to behave appropriately on the following trial.

ACKNOWLEDGEMENTS

This research was carried out during the senior author's tenure as a National Aeronautics and Space Administration predoctoral fellow under grant NSG(T)–137.

Reprints may be obtained from D. W. Massaro, Department of Psychology, University of California, San Diego, La Jolla, Calif. 92037.

REFERENCES

MASSARO, D. W. (1969). A three state Markov model for discrimination learning. *J. math. Psychol.* (in press).

MASSARO, D. W., HALPERN, J. & MOORE, J. W. (1968). Generalization effects in human discrimination learning with overt cue identification. *J. exp. Psychol.* **77**, 474–482.

POPPER, J. & ATKINSON, R. C. (1958). Discrimination learning in a verbal conditioning situation. *J. exp. Psychol.* **56**, 21–25.

APPENDIX

Presented below are the expressions for the conditional statistics of the form $P(A_{1,n+1}|T_{i,n+1}T_{j,n}A_{k,n}E_{l,n})$:

$$P(A_1|T_1T_1A_1E_1) = \frac{A+pU\{c_1+(1-c_1)p\}}{A+pU},$$

$$P(A_1|T_1T_1A_1E_2) = \frac{A\{cp+(1-c)\}+U(1-c_2)p^2}{A+pU},$$

$$\begin{aligned}
P(A_1|T_1T_1A_2E_1) &= \frac{Icp+(1-p)U\{c_2+(1-c_2)p\}}{I+(1-p)U},\\
P(A_1|T_1T_1A_2E_2) &= \frac{(1-p)U(1-c_1)p}{I+(1-p)U},\\
P(A_1|T_1T_2A_1E_1) &= \frac{U(1-c_1)p^2}{I+pU},\\
P(A_1|T_1T_2A_1E_2) &= \frac{Icp+pU\{c_2+(1-c_2)p\}}{I+pU},\\
P(A_1|T_1T_2A_2E_1) &= \frac{A\{cp+(1-c)\}+(1-p)U(1-c_2)p}{A+(1-p)U},\\
P(A_1|T_1T_2A_2E_2) &= \frac{A+(1-p)U\{c_1+(1-c_1)p\}}{A+(1-p)U},\\
P(A_1|T_2T_1A_1E_1) &= \frac{U(1-c_1)p^2}{A+pU},\\
P(A_1|T_2T_1A_1E_2) &= \frac{Acp+pU\{c_2+(1-c_2)p\}}{A+pU},\\
P(A_1|T_2T_1A_2E_1) &= \frac{I\{cp+(1-c)\}+(1-p)U(1-c_2)p}{I+(1-p)U},\\
P(A_1|T_2T_1A_2E_2) &= \frac{I+(1-p)U\{c_1+(1-c_1)p\}}{I+(1-p)U},\\
P(A_1|T_2T_2A_1E_1) &= \frac{I+pU\{c_1+(1-c_1)p\}}{I+pU},\\
P(A_1|T_2T_2A_1E_2) &= \frac{I\{cp+(1-c)\}+U(1-c_2)p^2}{I+pU},\\
P(A_1|T_2T_2A_2E_1) &= \frac{Acp+(1-p)U\{c_2+(1-c_2)p\}}{A+(1-p)U},\\
P(A_1|T_2T_2A_2E_2) &= \frac{(1-p)U(1-c_1)p}{A+(1-p)U}.
\end{aligned}$$

Vol. 21
Part 2
201–229

The British Journal of Mathematical and
Statistical Psychology

November
1968

STRUCTURAL PROPERTIES AND MEASUREMENT THEORY OF CERTAIN SETS ADMITTING A CONCATENATION OPERATION

By Walter Kristof
Educational Testing Service, Princeton, N.J.

This paper is a contribution to the theory of unidimensional scaling based on a concatenation operation defined for pairs of objects from certain continuous and ordered sets M. Inferences about structural properties of M, such as reflexivity and commutativity of the concatenation operation, are obtained by means of an entirely general non-metric treatment. It is shown that reflexivity holds either throughout or just for a single element and that commutativity of two different elements holds either generally or in no instance. These statements are independent. The study employs concatenation operations defined in an *ad hoc* way which formally and necessarily satisfy the reflexivity requirement and then also that of commutativity. By this means, the existence and essential uniqueness of scales of M with linear expressions for the various concatenation operations are established. Rules are derived for the way in which the parameters entering into these expressions can change with admissible scale transformations.

1. Introduction

This paper is intended to be a contribution to the theory underlying certain forms of unidimensional scaling. It is primarily concerned with inferences about the properties of sets M of objects which are encountered in psychophysics, e.g. sensations of loudness, brightness, and the like. Thus it belongs to the general area of measurement theory which has received growing interest since the publication of the Luce & Tukey (1964) paper on conjoint measurement. Current measurement theory concentrates on establishing existence and uniqueness theorems for numerical representations of collections of objects obeying various basic assumptions and so provides the theoretical bases for an increasing number of practical scaling methods. A review of achievements in this field, so far as psychology is concerned, is given in Tversky's (1967) recent paper. Another account has been given by Kristof (1967 *a*).

Some portions of the present paper were stimulated by the work of Pfanzagl (1959 *a*, *b*). He used the bisymmetry assumption adopted here; and he gave representation and uniqueness theorems for the case of a concatenation operation which is not necessarily reflexive and commutative, a topic dealt with in the concluding section of this paper. However, the present paper has a different main purpose: namely, an inquiry into such basic properties of M as reflexivity, commutativity and class membership of M (Def. 5) made entirely without the use of a previous scale of M. Apart from their theoretical interest, the main results in this area, R11 and R24, may be useful as checks for whether a domain can be scaled.

The paper has the following plan: some basic definitions and assumptions are given, then a study is made of the reflexivity of a concatenation operation and this is followed by an inquiry into commutativity with the aid of a specially defined formal reflexive operation of concatenation. Some further results on commutativity enable another operation to be defined which is necessarily commutative. On this basis, the problem of assigning numerical values to the elements of M, i.e. constructing scales without maintaining the reflexivity and commutativity requirements, can be formally reduced to the case of halving intervals. The existence and uniqueness problems of linear scales are solved for different sets of basic assumptions. Attention is given to the transformation properties of the numerical parameters involved.

The results are obtained in a strict way, but one which demands little mathematical knowledge. All necessary steps are given in detail unless reference can be made to earlier work. Consequently, a number of auxiliary theorems have been proved, especially about limits of infinite sequences, even though their counterparts in the domain of real numbers could be assumed as familiar. But a mathematically rigorous treatment requires convincing demonstrations, not just references to possible analogies.

2. BASIC ASSUMPTIONS

M is not necessarily supposed to be made up of numbers. Instead, its elements may be such things as pitch sensations, loudness or weight. What matters is not the nature of these elements but the possibility of defining relations between them in accord with the basic assumptions that will be described. This suffices to arrive at definite structural statements about M which can be verified empirically. From a mathematical viewpoint, an isomorphic mapping of M on to a certain set of real numbers (construction of a scale) means only a 'translation' of M and its structure into a more convenient language.

Let M be a set with typical elements $a, b, c, \ldots, x, y, z$. The sets which are of interest here may be defined by the following requirements.

Def. 1. (i) M is (simply) ordered by $>$, $<$ and $=$. Here $=$ expresses an equivalence relation between elements of M, whereas $>$ and $<$, respectively, denote a compatible ordering relation.

(ii) M possesses a maximal element, $x_{\max}$, and a minimal element, $x_{\min}$.

(iii) M is a continuous set. This means that every partitioning of M into a lower class A and an upper class Z with $M = A \cup Z$, $A \cap Z = 0$, defines a Dedekind cut, i.e. either A has a maximal element and Z has no minimal element, or A has no maximal element and Z has a minimal element. It follows that an ordered continuous set is necessarily dense, i.e. $x \in M$ and $y \in M$ with $x > y$, say, implies the existence of $z \in M$ with $x > z > y$.

M may be regarded as composed of ordered equivalence classes and the elements of M as representatives of such equivalence classes. The order relation may be given a probabilistic interpretation.

Def. 2. The sets considered here admit a concatenation operation which may be performed on any two elements of M. This means that $x \in M$ and $y \in M$ determine uniquely (in the sense of an equivalence relation) a third element $z \in M$ which may be written as $z = x * y$ or $z = x \circ y$, depending on the context.

Examples of such concatenation operations may be halving perceptual intervals which are defined by two given stimuli, or the construction of a stimulus resembling most two given stimuli. Time-order errors or tendencies to avoid extreme responses need not be ruled out.

Some basic assumptions, C1, C2, etc., are now formulated. Consequences, R1, R2, etc., will be drawn later, which are aimed at improving our insight into the structural properties of M.

*C*1. Each pair x, y of elements of M determines uniquely a third element $x * y \in M$ (existence).

This assumption is merely a repetition of Def. 2 and has been included for formal reasons. Further, for all elements:

*C*2. $x * y = y * x$ (commutativity, symmetry).

*C*3. $x * x = x$ (reflexivity, idempotence).

*C*4. $(x * y) * (u * v) = (x * u) * (y * v)$ (bisymmetry).

*C*5. $x \gtreqless x'$ iff $x * y \gtreqless x' * y$, respectively, and $y \gtreqless y'$ iff $x * y \gtreqless x * y'$ (monotonicity).

*C*6. $x * a > b$ together with $y * a < b$ implies the existence of $z \in M$ with $z * a = b$. Similarly, $a * u > b$, together with $a * v < b$, implies the existence of $w \in M$ with $a * w = b$.

It is easily seen with the help of C5 that the elements z and w in C6 are unique in the sense of an equivalence relation.

If C2 holds, then the second parts of C5 and C6 follow from the respective first parts of these assumptions. This is not the case if C2 is not required to apply throughout.

It has been shown previously (Kristof, 1967 *b*) that the validity of C1 to C6, together with the present definition of M, implies the existence of a real scale (numerical representation), $s(z)$, of M such that, for any $x \in M$ and $y \in M$, $x \gtreqless y$ iff $s(x) \gtreqless s(y)$, respectively, and $s(x * y) = \frac{1}{2}\{s(x) + s(y)\}$. Therefore, the operation $*$ may be aptly termed 'halving the interval between x and y'.

If, however, neither C2 nor C3 are expressly required, then the concatenation operation will be denoted by the symbol $\circ$ instead of $*$. But it will always be assumed that Def. 1, C1, C4, C5 and C6 apply, regardless of whether $*$ or $\circ$ is in use. The relations of the foregoing assumptions to those in Pfanzagl's (1959 *a*, *b*) work were outlined in the introduction, and there is a certain amount of overlap with his work in the results of Section 8.

3. A DISCUSSION OF REFLEXIVITY

It is the purpose of the present section to investigate whether definite statements concerning reflexivity of a concatenation operation $\circ$ can be made solely on the basis of Def. 1, C1, C4, C5 and C6, i.e. whether C3 is wholly independent of these assumptions. If it were possible to show dependence of C3 on the remaining assumptions (except C2), then the mathematical formulation of such dependence (if attainable) defines a necessary condition which data must satisfy in order to be compatible with the basic assumptions.

It will turn out that this can be shown. A sequence of conclusions will be presented leading step by step to the main theorem of this section, R11, which allows reflexivity to take on only one of two distinct alternatives.

*R*1. (1) Let a and b be two elements and consider the set A of all elements x satisfying $a \circ x < b$. Then $x \in A$ and $x' < x$ implies $x' \in A$. Furthermore, if A is not empty, and if it does not contain $x_{\max}$, then it has no maximal element, i.e. it is open to the right. (2) A completely analogous theorem holds for the set of all elements y satisfying $y \circ a < b$. Similarly, the following is true: (3) Consider the set B of all elements u satisfying $a \circ u > b$. Then $u \in B$ and $u' > u$ implies $u' \in B$. Furthermore, if B is not empty and if it does not contain $x_{\min}$, then it has no minimal element, i.e. it is open to the left. (4) A completely analogous theorem holds for the set of all elements v satisfying $v \circ a > b$.

Proof. R1 consists of four statements the proofs of which are all very similar. Therefore only (1) is proved in full detail. Suppose $x' < x$ and $x \in A$. C5 then yields $a \circ x' < a \circ x < b$, hence $x' \in A$. Now assume contrary to the assertion that A possesses a maximal element, x'', that is different from $x_{\max}$. Hence $a \circ x'' < b \leqslant a \circ x_{\max}$. C6 ensures the existence of an element z with $a \circ z = b$ and $x'' < z \leqslant x_{\max}$. The density of M guarantees that an element z' exists with $x'' < z' < z$ and, by C5, $a \circ z' < a \circ z = b$. This contradicts the maximality assumption of x'', hence the assertion is true. Q.E.D.

*R*2. (1) Let A be the set of all elements x satisfying $x \circ x < b$, b given. Then $x \in A$ and $x' < x$ implies $x' \in A$. Further, if A is not empty, and if it does not contain $x_{\max}$, then it has no maximal element. Similarly, the following is asserted: (2) Let B be the set of all elements y satisfying $y \circ y > b$, b given. Then $y \in B$ and $y' > y$ implies $y' \in B$. Further, if B is not empty and if it does not contain $x_{\min}$, then it has no minimal element.

Proof. It suffices to prove (1), the proof of (2) following similar lines. For the sake of clarity, the proof is broken down into several steps.

(i) Suppose $x' < x$ and $x \in A$. By C5, $x' \circ x' < x \circ x < b$, thus $x' \circ x' < b$ and $x' \in A$ by definition.

(ii) Assume, contrary to assertion, that A possesses a maximal element, x'', with $x'' < x_{\max}$. Thus $x'' \circ x'' < b$.

(iii) Consider the set S of all elements s satisfying $x'' \circ s < b$, x'' fixed. S is not empty since $x'' \in S$ by (ii). Further, S does not contain $x_{\max}$. Namely, suppose that $x_{\max} \in S$, i.e. $x'' \circ x_{\max} < b$. But $x_{\max} \circ x_{\max} \geqslant b$ due to x'' being a maximum, so $x'' \circ x_{\max} < b \leqslant x_{\max} \circ x_{\max}$. By C6, there is an element u with $u \circ x_{\max} = b$ and, by C5, $x'' < u \leqslant x_{\max}$. By C5, $u \circ u \leqslant u \circ x_{\max} = b$, thus $u \circ u \leq b$. If the sign = holds, then u' is chosen so that $x'' < u' < u$ which is possible due to the density of M. Hence, by C5, $u' \circ u' < b$. In any event, this contradicts the assertion that x'' is a maximum, hence S cannot contain $x_{\max}$. Therefore, according to R1, S has no maximal element.

(iv) Consequently, there is an element s' in S with $x'' < s'$ and $x'' \circ s' < b$. Now consider the set T of all elements t satisfying $t \circ s' < b$, s' fixed. Reasoning similar to (iii) shows that T has no maximal element and neither is it empty because $x'' \in T$.

(v) Then, however, there is an element t' with $x'' < t'$ and $t' \circ s' < b$. Let $\bar{x}$ be the 'smaller' one of the elements s' and t' or take $\bar{x} = s'$ if $s' = t'$. In each case, from C5 it follows that $\bar{x} \circ \bar{x} < b$ hence $\bar{x} \in A$. But $x'' < s'$ and $x'' < t'$ implies $x'' < \bar{x}$ which again contradicts the assumption of x'' being a maximum. Therefore A has no maximal element. Q.E.D.

*R*3. If, for b given, there are elements x and y with $x \circ x < b < y \circ y$, then there is a unique element z with $z \circ z = b$ and $x < z < y$.

Proof. The sets A and B of all elements x' and y' satisfying $x' \circ x' < b$ and $b < y' \circ y'$, respectively are, by assumption, not empty. By C5, they cannot contain $x_{\max}$ and $x_{\min}$, respectively. Then by R2, $x' \in A$ and $x'' < x'$ imply $x'' \in A$ and, on the other hand, $y \in B$ and $y'' > y'$ imply $y'' \in B$. Further, by R2, A has no maximal element and B has no minimal element. Now suppose there is no element z with $z \circ z = b$. Then this yields a partitioning $M = A \cup B$ of M, $A \cap Z = 0$, which is not a Dedekind cut and hence is a contradiction of part (iii) of Def. 1. Consequently, an element z with $z \circ z = b$ exists, and it follows from C5 that $x < z < y$. C5 also ensures $z' = z$ if $z' \circ z' = z \circ z$ which proves uniqueness. Q.E.D.

The following definition is now introduced.

Def. 3. $\sigma(z)$, $z \in M$, will be used to denote the element $z \circ z$ and it will be regarded as a function which maps M into (or on to) itself.

A quite general fixed point theorem is now derived.

*R*4. There is at least one element x with $\sigma(x) = x$, i.e. $\sigma(z)$ possesses at least one 'fixed point'.

Proof. Assume to the contrary that there is no fixed point. It follows that, in particular, $\sigma(x_{\min}) > x_{\min}$ and $\sigma(x_{\max}) < x_{\max}$. In general, there are elements u_0 and v_0 with $u_0 < v_0$ and $\sigma(u_0) > u_0$ and $\sigma(v_0) < v_0$. To establish R4 it will suffice to concentrate upon u_0. An infinite sequence of elements $u_0, u_1, u_2, \ldots, u_n, \ldots$ is defined as follows. Set $u_1 = \sigma(u_0)$ and $u_{n+1} = \sigma(u_n)$ for $n = 1, 2, \ldots$. The sequence of elements u_n is strictly increasing. Namely, $u_0 < u_1$ by assumption. Then, by C5, $u_1 = u_0 \circ u_0 < u_1 \circ u_1 = u_2$, hence $u_1 < u_2$. The same line of reasoning yields generally $u_n < u_{n+1}$. Since this sequence is bounded (certainly by $x_{\max}$) it will possess a supremum, $u = \sup u_n$, where u is the 'smallest' element 'larger' than all u_n (detailed proofs concerning the existence of limits, including suprema and infima, are contained in Kristof (1967 *a*)). It is $u > u_n$ for all n, hence, by C5, $u \circ u > u_n \circ u_n = u_{n+1}$ for all n. Therefore, it is certainly the case that $u \circ u \geqq u$, remembering the definition of u. But $u \circ u > u$ is impossible. Namely, the assumption to the contrary gives $u \circ u > u > u_n \circ u_n$ and R3 guarantees the existence of an element w with $u = w \circ w$ and $u > w > u_n$ for all n which contradicts u being the supremum. Hence $u \circ u = u$, i.e. $\sigma(u) = u$. Substituting x for u establishes R4. Q.E.D.

It should be noted that the proof of R4 did not require C4. Hence R4 is valid independently of C4 (and C2). C4 will be invoked now.

*R*5. If x and y are fixed points of $\sigma(z)$, $z \in M$, so are $x \circ y$ and $y \circ x$. Further, if $x < y$, then $x < x \circ y < y$ and $x < y \circ x < y$.

Proof. By C4, $(x \circ y) \circ (x \circ y) = (x \circ x) \circ (y \circ y) = x \circ y$. Similarly, $(y \circ x) \circ (y \circ x) = y \circ x$. Further, $x = x \circ x < x \circ y < y \circ y = y$ by C5 and $x < y \circ x < y$ analogously. Q.E.D.

This means that the set of fixed points of $\sigma(z)$, $z \in M$, is dense if there is more than one fixed point. However, density in M cannot be inferred from R5.

*R*6. (1) Suppose that $u_0 < \sigma(u_0) = u_1$. Let $u_{n+1} = \sigma(u_n)$ $(n = 1, 2, \ldots)$. Then $u = \sup u_n$ is the 'smallest' fixed point of $\sigma(z)$, $z \in M$, which is 'larger' than u_0. (2) Similarly, suppose that $v_0 > \sigma(v_0) = v_1$. Let $v_{n+1} = \sigma(v_n)$ $(n = 1, 2, \ldots)$. Then $v = \inf v_n$ is the 'largest' fixed point of $\sigma(z)$, $z \in M$, which is 'smaller' than v_0.

Proof. Only (1) is proved here, the proof of (2) being similar. It follows from the proof of R4 that the infinite sequence of elements u_n is strictly increasing and so possesses a supremum, $u = \sup u_n$, which is a fixed point of $\sigma(x)$. Now suppose there is another fixed point, $\bar{u}$, with $u_0 < \bar{u} < u$. From C5, $u_1 = u_0 \circ u_0 < \bar{u} \circ \bar{u} < u \circ u$, hence $u_1 < \bar{u} < u$. The same line of reasoning gives $u_n < \bar{u} < u$ for all n. Then, however, u cannot be $\sup u_n$. This contradiction proves the validity of R6. Q.E.D.

*R*7. Let x and y be fixed points of $\sigma(z)$, $z \in M$, with $x < y$. Let u_0 be any element with $x < u_0 < y$. Then u_0 is also a fixed point of $\sigma(z)$. In other words: All elements lying 'between' two fixed points of $\sigma(z)$ are fixed points themselves.

Proof. (i) If $u_0 = u_0 \circ u_0$, then there remains nothing to prove. Therefore let $u_0 < u_0 \circ u_0 = u_1$ (the other possibility $u_0 > u_0 \circ u_0$ would be handled similarly) and consider the strictly increasing sequence of elements $u_0, u_1, \ldots, u_n, \ldots$ with $u = \sup u_n$, this sequence being defined precisely as in the proof of R4. By R6, u is the 'smallest' fixed point of $\sigma(z)$ 'larger' than u_0, i.e. there is no other fixed point 'between' u_0 and u.

(ii) By assumption, $x = x \circ x < u_0 < u_0 \circ u_0$. Then, by R3, there is a unique element u'_1 with $u_0 = u'_1 \circ u'_1$ and $x < u'_1 < u_0$. Thus $x \circ x < u'_1 < u'_1 \circ u'_1$. Again by R3, there is a unique element u'_2 with $u'_1 = u'_2 \circ u'_2$ and $x < u'_2 < u'_1$. The same line of reasoning yields an infinite strictly decreasing sequence of elements, $\ldots < u'_n < \ldots < u'_2 < u'_1 < u_0$ with $u'_n = u'_{n+1} \circ u'_{n+1}$, bounded by x. Therefore this sequence possesses an infimum. $u' = \inf u'_n$, with $u' < u'_n$ for all n.

(iii) It will be proved that $u' = u' \circ u'$. To this end, suppose that $u' < u' \circ u'$. Then there must be an index m such that $u'_{m+1} \circ u'_{m+1} = u'_m < u' \circ u'$ from which we infer $u'_{m+1} < u'$ by C5. However, this contradicts u' being $\inf u'_n$. Now examine the possibility $u' > u' \circ u'$. It would then follow that $u' \circ u' < u' < u'_n = u'_{n+1} \circ u'_{n+1}$ for all n. By R3, there is an element v with $u' = v \circ v$ and

$u' < v < u'_{n+1}$ for all n. This contradicts u' being $\inf u'_n$. Thus the only alternative left is $u' = u' \circ u'$ which must therefore be true, i.e. $\sigma(u') = u'$.

(iv) u' is the 'largest' fixed point of $\sigma(z)$ which is 'smaller' than u_0. Namely, suppose there is another fixed point, u'', with $u' < u'' < u_0$. Then there must be an index k such that $u'_k < u''$. By C5 we obtain $u'_{k-1} = u'_k \circ u'_k < u'' \circ u'' = u''$, hence also $u'_{k-1} < u''$. After $k-1$ additional analogous steps we arrive at $u_0 < u''$ which contradicts the initial assumption $u'' < u_0$. Hence the assertion formulated in (iv) is correct.

(v) Therefore $u = \sup u_n$ and $u' = \inf u'_n$ limit an 'interval' which does not contain any fixed points of $\sigma(z)$ in its interior. However, u and u' are fixed points. This leads to a contradiction to the previous statement by virtue of R5. Consequently, $\sigma(u_0) = u_0$, i.e. R7 is valid. Q.E.D.

R8. If $\sigma(z)$, $z \in M$ possesses more than only one fixed point then there is a 'largest' and a 'smallest' one, i.e. the set of fixed points has a maximal and a minimal element.

Proof. (i) Only a detailed proof of the existence of a maximal fixed point will be given, the proof involving a minimal fixed point being similar. If $x_{\max}$ is a fixed point, then the assertion is trivially valid. It is assumed, therefore, that $x_{\max}$ is not a fixed point and that the set of fixed points has no maximal element, i.e. it is open to the right. Then the continuity of M requires the set of all elements 'greater' than all fixed points to possess a minimal element, u_0, in order that a Dedekind cut may result.

(ii) $u_0 \gtrless u_0 \circ u_0$ by definition. First, the possibility that $u_0 < u_0 \circ u_0$ is considered. An infinite sequence of elements u_n is defined by setting $u_1 = u_0 \circ u_0$ and $u_{n+1} = u_n \circ u_n$, i.e. $u_{n+1} = \sigma(u_n)$ generally. It has been already shown in the proof of R4 that a sequence of elements u_n as defined here is strictly increasing and possesses a supremum, $u = \sup u_n$, $u_0 < u$, for which $\sigma(u) = u$. This is incompatible with the definition of u_0 in view of R7 and so disproves $u_0 < u_0 \circ u_0$.

(iii) Suppose, on the other hand, that $u_0 > u_0 \circ u_0$. Let t be any fixed point of $\sigma(z)$, hence $t < u_0$. From C5, $t \circ t < u_0 \circ u_0$, hence $t < u_0 \circ u_0 < u_0$. The validity of this statement for all fixed points t contradicts u_0 being the 'smallest' element 'larger' than all fixed points as implicitly required by the continuity assumption used in (i).

(iv) Thus a contradiction is reached in either case. Consequently, the set of all elements 'greater' than all fixed points cannot possess a minimal element. Then, however, the set of all fixed points must have a maximal element in order that the continuity assumption may be satisfied. Q.E.D.

Combining R4, R7 and R8 it may be stated that the set of all fixed points of $\sigma(z) = z$, $z \in M$, is not empty and forms a 'closed interval' in M. This 'interval' may possibly reduce to a single element since this case has not been precluded.

Before further progress can be made it is necessary to prove R9.

R9. Let $x_0, x_1, \ldots, x_n, \ldots$ be an infinite and strictly monotone sequence with $x = \lim x_n$ being the supremum or infimum, whichever applies, $x_n \in M$.

For any fixed element c, $c \in M$, (1) $\lim(x_n \circ c) = (\lim x_n) \circ c$ and (2) $\lim(c \circ x_n) = c \circ (\lim x_n)$.

Proof. Only (1) is proved, the proof of (2) being similar. By C5, $x_n \gtrless x_{n+1}$ implies $x_n \circ c \gtrless x_{n+1} \circ c$, respectively. Therefore the infinite sequence of elements $x_n \circ c$ is in the same sense strictly monotone as the sequence of elements x_n and consequently it possesses a supremum or an infimum, whichever applies. It suffices to restrict ourselves to the case of a strictly increasing sequence because the other alternative can be treated in complete analogy. Thus $x > x_n$ for all n. By C5, $x \circ c > x_n \circ c$ for all n and so $x \circ c \geqslant \sup(x_n \circ c)$. Suppose that $x \circ c > \sup(x_n \circ c)$. Hence $x \circ c > \sup(x_n \circ c) > x_n \circ c$ for all n. By C6, there is an element x' with $\sup(x_n \circ c) = x' \circ c$ and $x > x' > x_n$, x' independent of n by C5. But this result contradicts x being $\sup x_n$. Therefore $x \circ c = \sup(x_n \circ c)$, which establishes R9. Q.E.D.

*R*10. If there are two different fixed points of $\sigma(z)$, $z \in M$, then all elements of M are such fixed points.

Proof. (i) According to R7 and R8, there is a ‘ closed interval ’ of fixed points $\sigma(z)$ in M. Let a be its minimal element and b its maximal element, $a < b$, and suppose that x_0 is an element with $x_0 < a$. Now choose an element c with $a < c \leq b$ which must be a fixed point according to R7.

(ii) An infinite sequence of elements $x_1, x_2, \ldots, x_n, \ldots$ is defined by $x_1 = x_0 \circ x_0$ and $x_{n+1} = x_n \circ x_n$. Here $x_1 = x_0$ is impossible because otherwise x_0 would be a fixed point in contradiction to (i). Also $x_1 < x_0$ cannot be true because $\inf x_n$ would be a fixed point ‘ smaller ’ than x_0 according to R6, thus contradicting (i). Consequently, $x_1 > x_0$ must hold and, by R6, $a = \sup x_n$.

(iii) A second infinite sequence of elements y_n is defined by $y_n = x_n \circ c$. (ii) implies also $x_n < x_{n+1}$ and so, by C5, $x_n \circ c < x_{n+1} \circ c$, i.e. $y_n < y_{n+1}$. Further, according to C4, $y_{n+1} = x_{n+1} \circ c = (x_n \circ x_n) \circ c = (x_n \circ x_n) \circ (c \circ c) = (x_n \circ c) \circ (x_n \circ c) = y_n \circ y_n$. Hence, $y_n < y_n \circ y_n$ and $y_n = x_n \circ c < c \circ c = c$, which implies $y_n < a$. Therefore, again by R6, $a = \sup y_n$. On the other hand, $y_n = x_n \circ c$ implies, by R9, $\sup y_n = (\sup x_n) \circ c$. Remembering (ii) we obtain $a = a \circ c$. However, by definition, $a = a \circ a$, thus $a \circ a = a \circ c$. C5 then requires $a = c$. This contradiction to the initial choice of c disproves the existence of an element x_0 with $x_0 < a$.

(iv) In complete analogy, there cannot be an element x_0 with $x_0 > b$ either. A detailed proof would parallel the one given previously and may therefore be omitted. Q.E.D.

R10 has the following obvious consequence. Precisely two cases are conceivable at this stage: (1) Either M has only one fixed point or (2) all elements of M are fixed points of $\sigma(z)$, $z \in M$. In other words: (1) Either just one element of M is reflexive or (2) all elements of M are reflexive. That at least one reflexive element exists has been established by R4. This latter theorem holds even regardless of C4.

It is now shown that the two cases just mentioned do indeed occur. (1.1) Let M be the closed segment of the real line [0, 1] and define

$x \circ y = \frac{1}{4}x + \frac{1}{2}y$. Obviously, this operation is an allowable definition of $\circ$. The only reflexive element is $x = 0$ and it is an extreme element.

(1.2) Let M be $[-1, 1]$ and define $\circ$ as before. Again the only reflexive element is $x = 0$, but this time it lies in the interior of M.

(2) Let M be $[0, 1]$ and define $x \circ y = \frac{1}{2}(x + y)$. Here all elements of M are reflexive.

It can easily be shown that, in each of the above cases, M is properly defined, and C1, C4, C5 and C6 are fulfilled.

The previous demonstration and R10 fully justify the following formulation of the main fixed point reflexivity theorem of this paper:

*R*11. If M satisfies the initial definition and C1, C4, C5 and C6 apply throughout, then precisely one of the following two cases occurs: (1) either M has exactly one reflexive element or (2) all elements of M are reflexive.

It should be remembered that none of the results of this section depends in any way upon C2, i.e. the commutativity of the operation $\circ$.

As a consequence of R11 it will be recognized that the general formulation of C3 in Section 2 is quite redundant. It suffices to say that there are two different reflexive elements in M, the only alternative being the coexistence of one reflexive and one non-reflexive element.

4. INTRODUCTION OF A NECESSARILY REFLEXIVE OPERATION

The next general aim is a study of the commutativity of the operation $\circ$, i.e. of C2. However, the argument will depend on the concatenation operation being reflexive without restrictions of any kind. Yet is has been shown in R11 that this requirement is not necessarily satisfied. This unsatisfactory situation can be remedied in the following way.

The starting-point is a set M and the assumptions C1, C4, C5 and C6 which will be met by an operation $\circ$. On this basis a new operation is introduced, denoted $\bullet$, by

Def. 4. $z = x \bullet y$ is written iff $z \circ z = x \circ y$.

It is now shown that this new operation satisfies the assumptions C1 and C3 to C6. In order to facilitate the distinction between assumptions referring to the operation $\circ$ and those involving the operation $\bullet$, a distinguishing accent (′) will be temporarily attached to the latter.

C1′. For any $x \in M$ and $y \in M$ with $x < y$, say, then by applying C5 twice $x \circ x < x \circ y < y \circ y$. According to R3 there is a unique element z with $z \circ z = x \circ y$. An analogous result obtains if $x > y$. Hence the unique element $z = x \bullet y$ generally exists. Q.E.D.

C3′. $x = x \bullet x$ is trivially valid for any $x \in M$. Q.E.D.

C4′. Let the elements x, y, u, v be given. Elements z, w, q, r, s, t are defined by $z \circ z = x \circ y$, $w \circ w = u \circ v$, $q \circ q = z \circ w$, $r \circ r = x \circ u$, $s \circ s = y \circ v$, $t \circ t = r \circ s$. C1′ ensures the meaningfulness of these definitions. The following may also be written: $z = x \bullet y$, $w = u \bullet v$, $q = z \bullet w$, $r = x \bullet u$, $s = y \bullet v$, $t = r \bullet s$.

Hence $q=(x \bullet y) \bullet (u \bullet v)$ and $t=(x \bullet u) \bullet (y \bullet v)$. By C4, $(x \circ y) \circ (u \circ v)=(x \circ u) \circ (y \circ v)$, thus $(z \circ z) \circ (w \circ w)=(r \circ r) \circ (s \circ s)$ by definition. Again by C4, $(z \circ w) \circ (z \circ w)=(r \circ s) \circ (r \circ s)$, thus $z \circ w=r \circ s$ by C5. Then by definition $q \circ q=t \circ t$ and so $q=t$ by C5. Therefore, by definition, $z \bullet w=r \bullet s$ and so $(x \bullet y) \bullet (u \bullet v)=(x \bullet u) \bullet (y \bullet v)$. This establishes C4′. Q.E.D.

C5′. Let $z \circ z=x \circ y$ and $z' \circ z'=x' \circ y$ for $x \in M$ and $y \in M$. Then $x \circ y \gtreqless x' \circ y$ iff $z \circ z \gtreqless z' \circ z'$ and, by C5, iff $z \gtreqless z'$, i.e. iff $x \bullet y \gtreqless x' \bullet y$, respectively. Thus C5 implies $x \gtreqless x'$ iff $x \bullet y \gtreqless x' \bullet y$, respectively. Analogously, $y \gtreqless y'$ iff $x \bullet y \gtreqless x \bullet y'$, respectively. This establishes C5′. Q.E.D.

C6′. Suppose that $x \bullet a>b$ and $y \bullet a<b$. By C5, $(x \bullet a) \circ (x \bullet a)>b \circ b>(y \bullet a) \circ (y \bullet a)$ which means precisely $x \circ a>b \circ b>y \circ a$. According to C6 there is a unique element z with $z \circ a=b \circ b$. However, we may write as well $b=z \bullet a$. Similarly, it is found that $a \bullet x>b$ and $a \bullet y<b$ implies the existence of a unique element w with $a \bullet w=b$. This establishes C6′. Q.E.D.

(Note that C3′ is always valid regardless of whether C3 holds.)

As to C2′; commutativity: For $x \in M$, $y \in M$ given, let $z \circ z=x \circ y$ and $w \circ w=y \circ x$. Then $x \circ y \gtreqless y \circ x$ iff $z \circ z \gtreqless w \circ w$ and, by C5, iff $z \gtreqless w$, i.e. iff $x \bullet y \gtreqless y \bullet x$, respectively. This means that $\circ$ is commutative in precisely those instances in which $\bullet$ is commutative. Further, in all non-commutative cases the signs $<$ and $>$ are preserved in going from $\bullet$ to $\circ$ and vice versa.

As a consequence, studying the commutativity of the operation $\circ$ is equivalent to studying the commutativity of the operation $\bullet$. This fact permits reflexivity of $\bullet$ to be assumed throughout. It should also be noted that the theorems R1, … *remain valid* if $\bullet$ is substituted for $\circ$ because in the derivation of these theorems commutativity was never presupposed.

5. A DISCUSSION OF COMMUTATIVITY

The general remarks at the beginning of Section 3 apply here as well if reflexivity is replaced by commutativity. It is the present aim to derive a theorem clearly determining the generality of $\bullet$ and so of $\circ$. The validity of C1, C4, C5 and C6 ensures the validity of C1′ and C3′ to C6′.

For formal reasons it is assumed that *there is a commutative pair of different elements u and v*, i.e. $u \bullet v=v \bullet u$. Also, without loss of generality, it is required that $u<v$. These assumptions will be used throughout.

*R*12. Consider the formulation $u \bullet y=x \bullet u$ with $u \leqslant x \leqslant v$. Hereby a unique function $y=\tau(x)$ is defined with $u \leqslant y \leqslant v$.

Proof. $u=\tau(u)$ is evident and $v=\tau(v)$ by assumption. Now suppose $u<x<v$. Then, by C5, $u \bullet u<x \bullet u<v \bullet u=u \bullet v$. According to C6′ there is an element y with $u \bullet y=x \bullet u$ and $u<y<v$. The uniqueness follows from the fact that $u \bullet y=u \bullet y'$ implies $y=y'$ by C5′. Q.E.D.

*R*13. $y=\tau(x)$ is strictly increasing and assumes every ‘ value ’ ‘ between ’ u and v, including these limits.

Proof. By C5′, $x \gtreqless x'$ implies $x \bullet u \gtreqless x' \bullet u$, hence $u \bullet y \gtreqless u \bullet y'$, and $y \gtreqless y'$, respectively. Further, it has been shown above that u and v are possible values of $\tau(x)$. Now choose an element y with $u<y<v$. Then, by C5′, $u \bullet u < u \bullet y < u \bullet v = v \bullet u$. According to C6′, there is an element x with $u \bullet y = x \bullet u$ and $u<x<v$. Q.E.D.

*R*14. Let $y' = \tau(x')$ and $y'' = \tau(x'')$. Then $\tau(x' \bullet x'') = \tau(x') \bullet \tau(x'')$ and $\tau(x)$ is an automorphism with respect to $\bullet$.

Proof. $u \bullet y' = x' \bullet u$ and $u \bullet y'' = x'' \bullet u$ implies $(u \bullet y') \bullet (u \bullet y'') = (x' \bullet u)(x'' \bullet u)$. C4′ then yields $(u \bullet u) \bullet (y' \bullet y'') = (x' \bullet x'') \bullet (u \bullet u)$ and, by C3′, $u \bullet (y' \bullet y'') = (x' \bullet x'') \bullet u$. In other words, $y' \bullet y'' = \tau(x' \bullet x'')$, therefore $\tau(x') \bullet \tau(x'') = \tau(x' \bullet x'')$. Taking R13 into account, it will be seen that $\tau(x)$ is an automorphism with respect to $\bullet$. Q.E.D.

In the sequel, it is required to show that $\tau(x)$ is the identity transformation, i.e. that all elements of the domain of $\tau(x)$ are fixed points of this transformation. It is already known that u and v are such fixed points. It is also clear by R14 that any element which can be obtained by successively applying the operation $\bullet$ to any finite number of fixed points is itself a fixed point.

Now the *tentative assumption* is introduced that the element x_0 with $u<x_0<v$ is not a fixed point of $\tau(x)$, i.e. $x_0 \neq \tau(x_0)$. Then an infinite sequence of elements $x_0, x_1, x_2, \ldots, x_n, \ldots$ is defined by $x_1 = \tau(x_0)$ and $x_{n+1} = \tau(x_n)$. It is evident by R12 and R13 that $u < x_n < v$ for all n. Further, this sequence is strictly increasing or strictly decreasing, depending on whether $x_0 < x_1$ or $x_0 > x_1$. Namely, suppose that $x_0 < x_1$. It follows from R13 that $\tau(x_0) < \tau(x_1)$, thus $x_1 < x_2$. The same line of reasoning gives generally $x_n < x_{n+1}$. Analogously, $x_0 > x_1$ leads to $x_n > x_{n+1}$ by the same argument. Since these sequences are evidently bounded in each event, they possess limits, $x = \lim x_n$, which will be either suprema or infima, depending on whether the respective infinite sequence is increasing or decreasing (detailed proofs concerning the existence of limits, including suprema and infima, are contained in Kristof (1967 *a*)). The assumption and definitions of this paragraph will be used in what follows.

*R*15. If $x = \lim x_n$, $x_1 = \tau(x_0)$, $x_{n+1} = \tau(x_n)$, x_0 given, then x is a fixed point of the automorphism $\tau(z)$. Further, x is the 'largest' ('smallest') fixed point which is 'smaller' ('larger') than x_0, whichever applies.

Proof. (i) In fuller detail

$$\begin{aligned} u \bullet x_1 &= x_0 \bullet u, \\ u \bullet x_2 &= x_1 \bullet u, \\ \vdots\quad & \quad\ \vdots \\ u \bullet x_{n+1} &= x_n \bullet u, \\ \vdots\quad & \quad\ \vdots \end{aligned}$$

By R9 and the paragraph preceding R15, the left sides and the right sides of these 'equations' have the limits $u \bullet x$ and $x \bullet u$, respectively. These limits must be equal because the corresponding sequences are pairwise identical. Thus $u \bullet x = x \bullet u$, i.e. $x = \tau(x)$ as has been asserted.

(ii) Suppose that $x_1 < x_0$, for example. Thus $x < x_0$. Let $\bar{x}$ be a fixed point with $x < \bar{x} < x_0$. Then, by C5′, $x \bullet u < \bar{x} \bullet u < x_0 \bullet u = u \bullet x_1$ and, since x and $\bar{x}$

are fixed points, $u \bullet x < u \bullet \bar{x} < u \bullet x_1$. Again by C5′, $x < \bar{x} < x_1$. The same line of reasoning yields generally $x < \bar{x} < x_n$ for all n which contradicts x being inf x_n. Therefore a fixed point as assumed above cannot exist. An analogous result is obtained by a similar argument for the case $x_0 < x_1$. Q.E.D.

R12 and R13 imply the existence of the inverse transformation $\tau^{-1}(z)$. It is an automorphism too. Namely, consider two elements $\tau^{-1}(x')$ and $\tau^{-1}(x'')$ belonging to the domain of $\tau(z)$. Then, by R14, $\tau\{\tau^{-1}(x') \bullet \tau^{-1}(x'')\} = \tau\{\tau^{-1}(x')\} \bullet \tau\{\tau^{-1}(x'')\} = x' \bullet x''$, hence $\tau^{-1}(x') \bullet \tau^{-1}(x'') = \tau^{-1}(x' \bullet x'')$.

For x_0 as assumed before, an infinite sequence of elements $x'_1 = \tau^{-1}(x_0)$ and $x'_{n+1} = \tau^{-1}(x'_n)$ is defined.

*R*16. If $x_{n+1} \lessgtr x_n$, then $x'_{n+1} \gtrless x'_n$, respectively, for all n.

Proof. R13 implies that $\tau^{-1}(z)$ is strictly increasing. It was $x_{n+1} \lessgtr x_n$ iff $x_1 \lessgtr x_0$, i.e. iff $\tau(x_0) \lessgtr x_0$ iff $\tau^{-1}\{\tau(x_0)\} \lessgtr \tau^{-1}(x_0)$, i.e. iff $x_0 \lessgtr x'_1$ iff $\tau^{-1}(x_0) \lessgtr \tau^{-1}(x'_1)$, i.e. iff $x'_1 \lessgtr x'_2$, i.e. iff $x'_2 \gtrless x'_1$. Continuing this chain by repeated application of τ^{-1} gives finally $x'_{n+1} \gtrless x'_n$. Q.E.D.

By analogy with the demonstration of R15 the following results can be obtained.

*R*17. If $x' = \lim x'_n$, $x'_1 = \tau^{-1}(x_0)$, $x'_{n+1} = \tau^{-1}(x_n)$, x_0 given, then x' is a fixed point of the automorphism $\tau^{-1}(z)$. Further, x' is the 'smallest' ('largest') fixed point which is 'larger' ('smaller') than x_0, whichever applies.

The proof of R17 parallels that of R15 completely. Q.E.D.

*R*18. The previously introduced element x_0 is a fixed point of the automorphism $\tau(z)$, i.e. $x_0 = \tau(x_0)$.

Proof. According to R16 and R17, $x \lessgtr x_0$ iff $x' \gtrless x_0$ with $x = \lim x_n$ and $x' = \lim x'_n$. Further, $z = \tau(z)$ iff $\tau^{-1}(z) = z$, i.e. the automorphisms $\tau(z)$ and $\tau^{-1}(z)$ have the same fixed points. Hence x and x' limit an interval with x_0 in its interior that contains no fixed points of the automorphism $\tau(z)$. But, according to R14, $\tau(x \bullet x') = \tau(x) \bullet \tau(x') = x \bullet x'$, thus $x \bullet x'$ is also a fixed point if x and x' are. It may further be inferred from $x \gtrless x'$, by using C5′, that $x = x \bullet x \lessgtr x \bullet x' \lessgtr x' \bullet x' = x'$, i.e. $x \lessgtr x \bullet x' \lessgtr x'$, respectively. So $x \bullet x'$ is a fixed point 'between' x and x' which contradicts the previous statement. Therefore $x_0 = \tau(x_0)$. Q.E.D.

This result can be reformulated as follows: If u and v with $u < v$ are commutative, then u and z are commutative for $u \leqslant z \leqslant v$. Analogously, v and z are commutative. This latter statement could be proved in the same way. It would first be shown that $v \bullet y = x \bullet v$ defines a unique function $y = \rho(x)$ and then the pattern of the previous proofs would be followed.

*R*19. Suppose $u \bullet v = v \bullet u$ with $u < v$. Consider *any* two elements, x and y, with $u \leqslant x \leqslant v$ and $u \leqslant y \leqslant v$. Then x and y are commutative.

Proof. Without loss of generality it may be assumed that $x < y$ and disregard the trivial case $x = y$. Thus $u \leqslant x < y \leqslant v$. It follows from the above paragraph that u and y are commutative and further that then y and x are commutative. Q.E.D.

In other words: If u and v are commutative, then these elements limit an interval such that any two elements of this interval, including the limiting elements, are commutative.

*R*20. Let V be the set of all elements v which satisfy $u \bullet v = v \bullet u$, u given. Then V possesses a maximal element and a minimal element, v' and v'', respectively.

Proof. Let x be an element with $x > u$ and suppose that x and u are not commutative. Then, by R19, x' and u cannot be commutative if $x' > x$. Now suppose that the set of all such elements x with $x > u$ has a 'smallest' element, x'', say. Thus $u \bullet x'' \gtrless x'' \bullet u$. On the other hand, there is certainly an element v satisfying $u \bullet v = v \bullet u$ with $v \geqslant u$ and $v < x''$ since $v = u$ evidently meets this condition. So V is not empty. The possibility $u \bullet x'' > x'' \bullet u$ is now considered; the other alternative would be treated analogously. From C5′, $u \bullet x'' > x'' \bullet u > v \bullet u = u \bullet v$. Then, by C6′, there exists an element w with $u \bullet w = x'' \bullet u$ and $x'' > w > v$. By C5′, it follows that $u \bullet w > w \bullet u$ in contradiction to the minimality of x''. Consequently, the set of all x as introduced above has no minimal element. Then the continuity of M requires V to have a maximal element. The existence of a minimal element of V can be proved similarly. Q.E.D.

*R*21. (1) If x and y are commutative, so also are $x \bullet a$ and $y \bullet a$ as well as $a \bullet x$ and $a \bullet y$ for every $a \in M$. (2) If $x \bullet a$ and $y \bullet a$ or $a \bullet x$ and $a \bullet y$ are commutative, so are x and y.

Proof. $x \bullet y = y \bullet x$ yields, by C5′, $(x \bullet y) \bullet (a \bullet a) = (y \bullet x) \bullet (a \bullet a)$, hence by C4′, $(x \bullet a) \bullet (y \bullet a) = (y \bullet a) \bullet (x \bullet a)$. This derivation can be reversed. The proof of the other alternative mentioned in R21 is analogous. Q.E.D.

*R*22. Let A be the set of all elements x with $u \leqslant x \leqslant a$. Let B be the set of all elements y with $b \leqslant y \leqslant u$. Suppose that any two elements of A are commutative and any two elements of B are commutative. Then any two elements of $A \cup B$ are commutative.

Proof. It is only necessary to show that $y \bullet x = x \bullet y$ if $u < x \leqslant a$ and $b \leqslant y < u$. To this end, an infinite sequence $y, y_1, y_2, \ldots, y_n, \ldots$ is defined by $y_1 = y \bullet x$ and $y_{n+1} = y_n \bullet x$. Evidently, $y < u < x$. This gives, with C3′ and C5′, $y = y \bullet y < y \bullet x < x \bullet x = x$; hence $y < y_1$ and $y_1 < x$. By an analogous argument, it follows that $y_n < y_{n+1}$ and $y_{n+1} < x$. Hence the infinite sequence of elements y_n is strictly increasing and bounded. Therefore $\sup y_n$ exists and, by R9 and $\sup y_{n+1} = \sup y_n$, it follows that $\sup y_n = (\sup s_n) \bullet x$. Using C3′, $(\sup y_n) \bullet (\sup y_n) = (\sup y_n) \bullet x$, hence, by C5′, $x = \sup y_n$. Consequently, there must be an index k such that $u < y_k < x$. By definition, $y_k \in A$ and so $y_k \bullet x = x \bullet y_k$. Since $y_k = y_{k-1} \bullet x$ and $x = x \bullet x$ it follows that $(y_{k-1} \bullet x) \bullet (x \bullet x) = (x \bullet x) \bullet (y_{k-1} \bullet x)$. Then, by R21, $y_{k-1} \bullet x = x \bullet y_{k-1}$. $k-1$ steps of this sort lead to the result $y \bullet x = x \bullet y$. Q.E.D.

*R*23. If there is one commutative pair of elements in M, $u \bullet v = v \bullet u$, $u < v$, then all pairs of elements are commutative.

Proof. (i) For u given, there is a maximal element, v', and a minimal element, v'', such that every pair of elements with members lying both either between v'' and u or between u and v', including the respective limits, is commutative. This follows from R19 and R20. Therefore all pairs of elements with members between v'' and v', including these limits, are commutative by R22.

(ii) It is known that $u < v'$ because of $u < v$. Now suppose that $v' < x_{\max}$. Consider an element r with $v' < r \leqslant x_{\max}$. By R21 and R19, all pairs of elements with members lying between $u \bullet r$ and $v' \bullet r$ are commutative, these limits included. For these limits: $u = u \bullet u < u \bullet r$, thus $u < u \bullet r$; and $v' < v' \bullet r$. Further, by C5′, $u \bullet r < v' \bullet r$.

(iii) It is certainly possible to choose r such that $u \bullet r \leqslant v'$, even though $v' < r$. For suppose that an element r' is chosen instead of such an r, then $u \bullet r' > v' = v' \bullet v' > v' \bullet u = u \bullet v'$, by C3′, C5′, $v' > u$ and the commutativity of u and v'. So $u \bullet r' > v' > u \bullet v'$. According to C6′, there is then an element r with $v' = u \bullet r$ and $r' > r > v'$. Hence the condition $u \bullet r \leqslant v'$ can always be satisfied by a proper choice of r.

(iv) u and v' as well as $u \bullet r$ and $v' \bullet r$ border ‘ intervals ’ $[u, v']$ and $[u \bullet r, v' \bullet r]$, respectively, such that every pair of elements with both members contained in either of these intervals is commutative. The same holds for the interval $[v', v' \bullet r]$ since it is a proper or improper sub-interval of $[u \bullet r, v' \bullet r]$. But the intervals $[u, v']$ and $[v', v' \bullet r]$ share the element v' and are adjacent, and hence R22 applies. It follows that all pairs of elements with members between u and $v' \bullet r$, including these limits, are commutative. But $v' < v' \bullet r$, as stated in (ii). This contradicts the maximality property of v' postulated in (i). So the assumption $v' < x_{\max}$ must be false, i.e. $v' = x_{\max}$.

(v) It can be shown by an analogous reasoning that $v'' = x_{\min}$. Then according to (i) any two elements of M are commutative. Q.E.D.

R23 has the following obvious consequence. Precisely two cases are conceivable at this stage: (1) either there is no pair of commutative elements u and v, $u \neq v$, or (2) any two elements of M are commutative. It should be remembered that this statement remains correct regardless of whether it refers to an operation $\bullet$ or $\circ$, as has been generally stressed in the previous section.

It will now be shown that the two cases mentioned above do indeed occur.

(1) Let M be the closed segment of the real line $[0, 1]$ and define $x \circ y = \frac{1}{4}x + \frac{1}{2}y$. There is no commutative pair of different elements.

(2) Define M as above and set $x \circ y = \frac{1}{4}x + \frac{1}{4}y$. Here, any two elements are commutative.

It may be readily determined that in each of these two cases M is properly defined and C1, C4, C5 and C6 are fulfilled.

The previous demonstration and R23 fully justify the following formulation of the main commutativity theorem of this paper.

*R*24. If M satisfies the initial definition and C1, C4, C5 and C6 generally apply, then precisely one of the following two cases occurs: (1) either M contains no pair of different commutative elements or (2) every pair of elements of M is commutative.

It will now be demonstrated that R11 and R24, the main results, are independent in the sense that each combination of the cases allowed by these theorems does occur. Let M be the closed segment of the real line [0, 1]. The following fourfold table is self-explanatory.

	Commutativity	
Reflexivity	Always	In no instance
Always	$x \circ y = \frac{1}{2}x + \frac{1}{2}y$	$x \circ y = \frac{1}{4}x + \frac{3}{4}y$
In just one instance	$x \circ y = \frac{1}{4}x + \frac{1}{4}y$	$x \circ y = \frac{1}{4}x + \frac{1}{2}y$

As a consequence of R24 it will be seen that the general formulation of C3 given in Section 2 is quite redundant. It suffices to say that there is one commutative pair of different elements in M, the only alternative being the absence of any such commutative pair.

6. Some Further Results on Commutativity

In this section some additional results concerning commutativity are given. They may be viewed as necessary conditions which commutativity must obey in order to be consistent with the definition of M, C1 and C4 to C6. It is believed that these conditions may be empirically checked for a set of data.

The derivations will again be made in terms of the operation $\bullet$. However, it is known that the results on commutativity remain valid when $\circ$ is eventually substituted for $\bullet$. Only the case permitted by R24 in which no two different elements of M are actually commutative will be considered.

The following theorem is first proved:

R25. Suppose that $a \bullet x < x \bullet a$ and $a \bullet z > z \bullet a$; all three elements a, x, z being pairwise different. Then a necessarily lies ‘ between ’ x and z.

Proof. (i) It is assumed that $x < z$, since the other alternative may be treated similarly. Further, let $a < x$, contrary to the assertion (again, the case $z < a$ can be treated analogously). Hence $a < x < z$. By C5′, the initial assumptions can be combined to give $a \bullet x < x \bullet a < z \bullet a < a \bullet z$. Then, according to C6′, there is an element r such that $a \bullet r = x \bullet a$ and $x < r < z$. Likewise, there is an element t with $a \bullet t = z \bullet a$ and $x < t < z$. By C5′ $x < z$ implies $x \bullet a < z \bullet a$ and so $a \bullet r < a \bullet t$, therefore $r < t$. Thus $x < r < t < z$.

(ii) Every element y together with the relation $x \leqslant y \leqslant z$ determines uniquely an element s with $r \leqslant s \leqslant t$ when $a \bullet s = y \bullet a$ is required. The uniqueness follows from setting $a \bullet s = a \bullet s'$, which yields $s = s'$ by C5. As to existence, $s = r$ for

$y=x$ and $s=t$ for $y=z$ has been already established in (i). Therefore it is assumed that $x<y<z$. Then, by C5′ and (i), it follows that $a \bullet r = x \bullet a < y \bullet a < z \bullet a = a \bullet t$. Hence, by C6′, there is an element s with $a \bullet s = y \bullet a$ and $r<s<t$.

(iii) An infinite sequence of elements, $s_1, s_2, \ldots, s_n, \ldots$, is defined by

$$\begin{aligned} a \bullet r &= x \bullet a, \\ a \bullet s_1 &= r \bullet a, \\ a \bullet s_2 &= s_1 \bullet a, \\ &\vdots \\ a \bullet s_{n+1} &= s_n \bullet a, \\ &\vdots \end{aligned}$$

This definition is meaningful. Namely, all elements s_n do exist according to (i) and (ii). Further, by (ii), generally $r \leqslant s_n \leqslant t$ for all n. In addition, by C5′, $x<r$ implies $r<s_1<s_2<\ldots<s_n<\ldots$. Therefore the sequence of elements s_n is strictly increasing and bounded by t. Thus it possesses a supremum, $\bar{s}$, with $r<\bar{s}\leqslant t$.

(iv) It follows from the above ‘ equations ’ and R9 that $a \bullet \bar{s} = \bar{s} \bullet a$. Hence a and $\bar{s}$ are a commutative pair of elements with $a<x<r<\bar{s}$, i.e. $a<\bar{s}$. In this case, however, R24 requires any two elements to be commutative which contradicts the initial assumptions. This proves R25. Q.E.D.

R26. (1) Suppose that $a \bullet x < x \bullet a$ and $a \bullet z < z \bullet a$, all three elements a, x, z being different. Then both x and z are either ‘ greater ’ or ‘ smaller ’ than a. (2) The same holds when $>$ applies throughout.

Proof. Only (1) will be treated, the proof of (2) being similar.

(i) It is assumed without loss of generality that $x<z$ and $x<a<z$ contrary to the assertion. Combining corresponding sides of the above ‘ inequalities ’ by the operation $\bullet$ and taking C3′, C4′ and C5′ into account gives $a \bullet (x \bullet z) < (x \bullet z) \bullet a$. It follows that $x \bullet z \neq a$. Otherwise, namely, the contradiction $a \bullet a < a \bullet a$ would result.

(ii) Thus either $x \bullet z < a$ or $x \bullet z > a$ (the second case would be treated analogously). Suppose the first alternative holds. Then $x \bullet z < a < z = z \bullet z$ by (i) and C3′. It follows by C6′ that there is an element y with $a = y \bullet z$ and $x<y<z$. However, $y=a$ is impossible because otherwise $a = a \bullet a = a \bullet z$ by C3′ and so $a=z$ by C5′ in contradiction to (i). Thus $y \neq a$. Then, according to (i), either x and y or z and y lie on the same side of a. It follows from R25 and the initial ‘ inequalities ’ that in each instance $a \bullet y < y \bullet a$. But it is also known that $a \bullet z < z \bullet a$. Combining corresponding sides of these ‘ inequalities ’ by the operation $\bullet$ and applying C3′, C4′ and C5′ yields $a \bullet (y \bullet z) < (y \bullet z) \bullet a$ which reduces to $a \bullet a < a \bullet a$ in view of $a = y \bullet z$. This contradiction establishes R26. Q.E.D.

R27. (1) If $a<x$ and $a \bullet x<x \bullet a$, then $b<z$ implies $b \bullet z<z \bullet b$ and vice versa. (2) If $a<x$ and $a \bullet x>x \bullet a$, then $b<z$ implies $b \bullet z>z \bullet b$ and vice versa.

Proof. Only (1) will be proved, the proof of (2) being similar.

(i) Let $b \leqslant a$, hence $b \leqslant a<x$, i.e. both b and a are on the same side of x. Then, by R25 together with R26, $b \bullet x<x \bullet b$. Since x and z are both on the same side of b, then analogously $b \bullet z<z \bullet b$.

(ii) Let $a<b$ and suppose contrary to the assertion that $b \bullet z>z \bullet b$ (the sign $=$ cannot apply in view of R24), so $z \bullet b<b \bullet z$. Since a and b lie on the same side of z, then, by R25 and R26, $z \bullet a<a \bullet z$. Thus $a \bullet z>z \bullet a$. But x and z lie on the same side of a, which again leads to the result $a \bullet x>x \bullet a$ in contradiction to the assumption.

(iii) Vice versa, it is clear by the same arguments that $b \bullet z>z \bullet b$ is compatible only with $b<z$. Q.E.D.

The cases not excluded by R25 and R26 do indeed occur. For example, let $M=[-1, +1]$, i.e. the closed segment of the real line bordered by -1 and $+1$. Consider the operation $u \circ v=\frac{1}{4}u+\frac{3}{4}v$, $u \in M$, $v \in M$. It can easily be seen that this definition is consistent with the initial axioms.

Now consider R25. Put $a=0$, $x=-1$, $z=1$. Thus $a \circ x=-\frac{3}{4}$, $x \circ a=-\frac{1}{4}$, $a \circ z=\frac{3}{4}$, $z \circ a=\frac{1}{4}$. Clearly, $a \circ x<x \circ a$, $a \circ z>z \circ a$, $x<a<z$.

Consider R26. Take $a=0$, $x=\frac{1}{2}$, $z=1$. Then $a \circ x=\frac{3}{8}$, $x \circ a=\frac{1}{8}$, $a \circ z=\frac{3}{4}$, $z \circ a=\frac{1}{4}$. Then $a \circ x>x \circ a$, $a \circ z>z \circ a$, $a<x<z$.

Due to R27, the actual choice of a and x is immaterial as long as the order relation of a and x is preserved if relations of the sort $a \circ x \gtrless x \circ a$ are to be studied.

In the above examples, use was made of the fact that the results remain valid, as far as commutativity is concerned, if $\circ$ is substituted for $\bullet$.

The results of this section are summarized in the following self-explanatory fourfold table.

	Inequality signs in $a \circ x \lessgtr x \circ a$ and $a \circ z \lessgtr z \circ a$ are	
Inequality signs in $a \lessgtr x$ and $a \lessgtr z$ are	Same	Different
Same	Possible	Impossible
Different	Impossible	Possible

7. INTRODUCTION OF A NECESSARILY COMMUTATIVE OPERATION

The operation $\circ$ did not require the general validity of C2 and C3. However, an operation which was also reflexive was obtained by defining $\bullet$. This operation enabled a study to be made of the commutativity which led to the results of the last two sections.

It is now intended to introduce an operation which generally provides commutativity. This operation then satisfies all assumptions C1 to C6. It will be shown later that it is possible to assign on this basis numerical values to the elements of M such that an isomorphic map of M results.

The following theorem is first considered:

*R*28. Let u_n and v_n, respectively, denote the general elements of two infinite and strictly monotone sequences of elements of M with limits u and v, respectively. Assume further that the infinite sequence with the general element $u_n \bullet v_n$ is strictly monotone. Then $\lim(u_n \bullet v_n) = (\lim u_n) \bullet (\lim v_n)$.

Proof. (i) Consider first the case in which the initial sequences are both increasing (the case of two decreasing sequences would be handled similarly). Thus $u_n < u_{n+1}$ and $v_n < v_{n+1}$ for all n. Hence, by C5′, $u_n \bullet v_n < u_{n+1} \bullet v_{n+1}$ without invoking an additional assumption. Therefore the sequence with the general element $u_n \bullet v_n$ is strictly increasing. Then $u_n < u$ and $v_n < v$ for all n; hence, by C5′, $u_n \bullet v_n < u \bullet v$ for all n. It follows that $w = \sup(u_n \bullet v_n)$ exists and $w \leqslant u \bullet v$. On the other hand, an index k may be chosen and consideration given to the two infinite and strictly monotone sequences $u_k \bullet v_k$, $u_{k+1} \bullet v_{k+1}$, $u_{k+2} \bullet v_{k+2}, \dots$ and $u_k \bullet v_k$, $u_k \bullet v_{k+1}$, $u_k \bullet v_{k+2}, \dots$ with limits w and $u_k \bullet v$, by definition and by R9. Since, by C5′, $u_{k+1} \bullet v_{k+1} > u_k \bullet v_{k+1}$, $u_{k+2} \bullet v_{k+2} > u_k \bullet v_{k+2}$, and so on, it follows that $w \geqslant u_k \bullet v$ for all k. Hence $w \geqslant \sup\limits_{k \to \infty} (u_k \bullet v) = u \bullet v$ by R9. Therefore $w \leqslant u \bullet v$ and $w \geqslant u \bullet v$, hence $w = u \bullet v$.

(ii) Now consider the case in which one of the initial sequences increases and the other decreases, i.e. $u_n < u_{n+1}$ and $v_{n+1} < v_n$ for all n. It is further assumed that $u_n \bullet v_n < u_{n+1} \bullet v_{n+1}$ for all n (the other case would be treated analogously). Since, by C5′, $u_n \bullet v_n < u \bullet v_1$ for all n, $w = \sup(u_n \bullet v_n)$ exists. Also, by C5′, $u_n \bullet v < u_{n+1} \bullet v_{n+1} < u \bullet v_n$ for all n which yields $\sup(u_n \bullet v) \leqslant \sup(u_{n+1} \bullet v_{n+1}) \leqslant \inf(u \bullet v_n)$, i.e. by R9, $u \bullet v \leqslant w \leqslant u \bullet v$, thus $w = u \bullet v$. Q.E.D.

Next, a distinction is made between two exhaustive and mutually exclusive classes of sets M for which commutativity of the operation $\bullet$ (and consequently of the operation $\circ$) does not generally hold.

Def. 5. A set M is said to be of class I when, for all x and y with $x < y$, the relation $x \bullet y < y \bullet x$ holds. M shall be called of class II if, for all x and y with $x < y$, the relation $x \bullet y > y \bullet x$ is satisfied, x and y being elements of M.

It follows from R27 that this definition is meaningful and unambiguous. The class membership of a set M can be identified by examining any two different elements of it. It should be noted that class membership remains unchanged if $<$ is generally replaced by $>$ and vice versa. Therefore it constitutes an invariant with respect to notation and so expresses a structural property of M.

The definitions of a necessarily commutative operation will be chosen slightly differently for the two classes of sets M. Class I will be considered first.

Def. 6. Let M be of class I. Suppose x and y are different elements of M, i.e. $x \gtrless y$. Then $x \bullet y \gtrless y \bullet x$, respectively, according to Def. 5. Two infinite sequences with elements x_n and y_n are defined recursively by

$$\begin{array}{lcl} x_1 = x \bullet y, & \text{and} & y_1 = y \bullet x, \\ x_2 = x_1 \bullet y_1, & & y_2 = y_1 \bullet x_1, \\ \vdots & & \vdots \\ x_{n+1} = x_n \bullet y_n, & & y_{n+1} = y_n \bullet x_n, \\ \vdots & & \vdots \end{array}$$

Suppose that $x < y$. Then, by C3′ and C5′, $x = x \bullet x < x \bullet y < y \bullet y = y$, thus $x < x_1 < y$. Similarly, $x < y_1 < y$ and by assumption (class I), $x < x_1 < y_1 < y$. Further, $x_1 = x_1 \bullet x_1 < x_1 \bullet y_1 = x_2$, hence $x_1 < x_2$. Similarly, $y_2 < y_1$. It follows from $x_1 < y_1$ and by the definition of class I that $x_1 \bullet y_1 < y_1 \bullet x_1$, hence $x_2 < y_2$. The same line of reasoning shows that generally $x < x_1 < x_2 < \ldots < x_n < \ldots$ and $\ldots < y_n < \ldots < y_2 < y_1 < y$ with $x_n < y_n$ for all n. Hence $\bar{x} = \sup x_n$ and $\bar{y} = \inf y_n$ exist. Considering the above 'equations' and R28, we obtain $\bar{x} = \bar{x} \bullet \bar{y}$ as well as $\bar{y} = \bar{y} \bullet \bar{x}$. Then, by C3′ and C5′, $\bar{x} = \bar{x} \bullet \bar{x} = \bar{x} \bullet \bar{y}$ and so $\bar{x} = \bar{y}$. It is evident that $x < \bar{x} < y$. Since $\bar{x} = \bar{y}$, this element is independent of the order relation in $x \gtrless y$, i.e. it is associated with the (unordered) pair of elements x, y. This element is indicated simply as $x \triangle y$. It is obviously uniquely determined by x and y and lies between x and y.

If $x = y$ is formally inserted in the above 'equations', then $\lim x_n = x$ by C3′. In this sense, $x \triangle x = x$ for all x.

Thus an operation $\triangle$ has been defined. It will be shown that this operation satisfies the assumptions C1 to C6. For formal reasons, these will now be referred to as C1″ to C6″ if they involve the derived operation $\triangle$.

It follows at once that

C1″. Any two elements x and y of M determine uniquely an element $z = x \triangle y$ of M.

C2″. $x \triangle y = y \triangle x$.

C3″. $x \triangle x = x$.

The verification of the remaining assumptions (axioms) is not so immediate. The following derivations refer to sets M of class I unless stated otherwise. It is helpful to prove first

*R*29. Consider four elements, x, y, u, v. Then $(x \triangle y) \bullet (u \triangle v) = (x \bullet u) \triangle (y \bullet v)$.

Proof. Put $w = x \bullet u$ and $z = y \bullet v$ and define

$$\begin{array}{ll} w_1 = w \bullet z, & z_1 = z \bullet w, \\ w_2 = w_1 \bullet z_1, & z_2 = z_1 \bullet w_1, \\ \vdots & \vdots \\ w_{n+1} = w_n \bullet z_n, & z_{n+1} = z_n \bullet w_n, \\ \vdots & \vdots \end{array}$$

By Def. 6, $\lim w_n = \lim z_n = w \vartriangle z = (x \bullet u) \vartriangle (y \bullet v)$. But the above 'equations' may be rewritten as follows, using C4′.

$$\begin{aligned} w_1 &= (x \bullet u) \bullet (y \bullet v) & \qquad z_1 &= (y \bullet v) \bullet (x \bullet u) \\ &= (x \bullet y) \bullet (u \bullet v) & &= (y \bullet x) \bullet (v \bullet u) \\ &= x_1 \bullet u_1. & &= y_1 \bullet v_1. \\ w_2 &= (x_1 \bullet u_1) \bullet (y_1 \bullet v_1) & z_2 &= (y_1 \bullet v_1) \bullet (x_1 \bullet u_1) \\ &= (x_1 \bullet y_1) \bullet (u_1 \bullet v_1) & &= (y_1 \bullet x_1) \bullet (v_1 \bullet u_1) \\ &= x_2 \bullet u_2. & &= y_2 \bullet v_2. \end{aligned}$$

Generally, for all n,

$$w_n = x_n \bullet u_n, \qquad z_n = y_n \bullet v_n.$$

In these reformulations the definitions of x_n, u_n, y_n and v_n follow the notation used in Def. 6 and are therefore not believed to require further explanation. Now it follows from R28 that $\lim w_n = (\lim x_n) \bullet (\lim u_n)$, i.e. remembering Def. 6, $(x \bullet u) \vartriangle (y \bullet v) = (x \vartriangle y) \bullet (u \vartriangle v)$ as has been asserted. Q.E.D.

Now the following is proved:

C5″. $x' \lesseqgtr x$ iff $x' \vartriangle y \lesseqgtr x \vartriangle y$ for all y. Also, $y' \lesseqgtr y$ iff $x \vartriangle y' \lesseqgtr x \vartriangle y$ for all x.

Proof. It suffices to prove the first statement in view of C2″.

(i) We define

$$\begin{aligned} x'_1 &= x' \bullet y, & \qquad y'_1 &= y \bullet x', \\ x'_2 &= x'_1 \bullet y'_1, & y'_2 &= y'_1 \bullet x'_1, \\ &\vdots & &\vdots \\ x'_{n+1} &= x'_n \bullet y'_n, & y'_{n+1} &= y'_n \bullet x'_n, \\ &\vdots & &\vdots \end{aligned}$$

Correspondingly,

$$\begin{aligned} x_1 &= x \bullet y, & \qquad y_1 &= y \bullet x, \\ x_2 &= x_1 \bullet y_1, & y_2 &= y_1 \bullet x_1, \\ &\vdots & &\vdots \\ x_{n+1} &= x_n \bullet y_n, & y_{n+1} &= y_n \bullet x_n, \\ &\vdots & &\vdots \end{aligned}$$

By Def. 6, $x' \vartriangle y = \lim x'_n$ and $x \vartriangle y = \lim x_n$. Using C3′, $x' \lesseqgtr x$ implies $x'_n \lesseqgtr x_n$ and $y'_n \lesseqgtr y_n$, respectively. It is evident that $x' = x$ implies $x' \vartriangle y = x \vartriangle y$.

(ii) Consider the case $x' < x$. It follows only that $\lim x'_n \leqslant \lim x_n$, i.e. $x' \vartriangle y \leqslant x \vartriangle y$. Now suppose that $x' \vartriangle y = x \vartriangle y$. Let x'' be any element with $x' < x'' < x$. We obtain $x' \vartriangle y \leqslant x'' \vartriangle y \leqslant x \vartriangle y = x' \vartriangle y$, therefore $x' \vartriangle y = x'' \vartriangle y = x \vartriangle y$.

(iii) Choose an element x'' with $x' < x'' < x$ and define an infinite sequence of elements $x''_1 = x'' \bullet y$, $x''_2 = x'' \bullet x''_1$; and generally $x''_{n+1} = x'' \bullet x''_n \cdots$ It is possible to arrange that $x'' \neq y$ (just to avoid the necessity to discuss this case separately). From C5′ it is found that this sequence is strictly monotone and bounded by x''. So it has a limit. R9 gives $\lim x''_n = x'' \bullet \lim x''_n$. C3′ and C5′ yield $x'' = \lim x''_n$. Therefore, by taking a finite index k large enough, it is possible to have $x' < x''_k < x$. Then, by (ii), $x''_k \triangle y = x \triangle y$.

(iv) Remembering the definitions in (iii), R29 and C3″,

$$\begin{aligned}
x''_1 \triangle y &= (x'' \bullet y) \triangle y = (x'' \triangle y) \bullet y,\\
x''_2 \triangle y &= (x'' \bullet x''_1) \triangle y = (x'' \triangle y) \bullet (x''_1 \triangle y),\\
x''_3 \triangle y &= (x'' \bullet x''_2) \triangle y = (x'' \triangle y) \bullet (x''_2 \triangle y),\\
&\vdots\\
x''_k \triangle y &= (x'' \bullet x''_{k-1}) \triangle y = (x'' \triangle y) \bullet (x''_{k-1} \triangle y).
\end{aligned}$$

But $x''_k \triangle y = x \triangle y$ by (iii) and $x'' \triangle y = x \triangle y$ by (ii). The last 'equation' yields then, using C3′ and C5′, $x \triangle y = x''_{k-1} \triangle y$. Introducing this into the 'equation' before the last, it is found that $x \triangle y = x''_{k-2} \triangle y$. After a finite number of analogous backward steps the result $x \triangle y = y$ is obtained. However, it was seen in Def. 6 that $x \triangle y$ must lie between x and y if $x \neq y$ (the case $x = y$ is trivial in view of this requirement and C3″). This contradiction proves $x' \triangle y < x \triangle y$.

(v) Similarly, it may be shown that $x' > x$ implies $x' \triangle y > x \triangle y$. But if $x' \lesseqgtr x$ implies $x' \triangle y \lesseqgtr x \triangle y$, respectively, then the corresponding reverse implications are obviously also correct. Q.E.D.

C6″. If, for a and b given, $x \triangle a < b < y \triangle a$, then there is a unique element z with $z \triangle a = b$. Similarly, $a \triangle u < b < a \triangle v$ implies the existence of a unique element w with $a \triangle w = b$.

Proof. In view of C2″, it suffices to prove only the first part.

(i) Let A be the set of all elements x satisfying $x \triangle a < b$ and let B be the set of all elements y satisfying $y \triangle a > b$. Then, by C5″, $x' < x$ and $x \in A$ implies $x' \in A$. Similarly, $y' > y$ and $y \in B$ implies $y' \in B$. The sets A and B are not empty by assumption. Suppose that an element z with $z \triangle a = b$ does not exist. Then, due to the continuity of M, A has either a 'largest' element and B has no 'smallest' or A has no 'largest' element and B has a 'smallest'. Suppose the first case holds (the second would be treated similarly). Let x'' be the 'largest' element of A, i.e. $x'' \triangle a < b$ and $y \triangle a > b$ for all y with $y > x''$.

(ii) An infinite sequence of elements is defined by $x''_1 = x'' \bullet y$, $x''_2 = x'' \bullet x''_1$, …, $x''_{n+1} = x'' \bullet x''_n$, …. C5′ shows that this sequence is strictly decreasing and bounded by x'', so it has a limit. R9 gives $\lim x''_n = x'' \bullet \lim x''_n$. C3′ and C5′ yield $x'' = \lim x''_n$. Since $x''_n > x''$ for all n, it follows that $x''_n \in B$ and so $x''_n \triangle a > b$ for all n.

(iii) Consider the infinite sequence of elements $x''_n \triangle a$. From R29 and C3″,

$$\begin{aligned}
x''_1 \triangle a &= (x'' \bullet y) \triangle a = (x'' \triangle a) \bullet (y \triangle a),\\
x''_2 \triangle a &= (x'' \bullet x''_1) \triangle a = (x'' \triangle a) \bullet (x''_1 \triangle a),\\
x''_3 \triangle a &= (x'' \bullet x''_2) \triangle a = (x'' \triangle a) \bullet (x''_2 \triangle a),\\
&\vdots \qquad\qquad\qquad \vdots\\
x''_{n+1} \triangle a &= (x'' \bullet x''_n) \triangle a = (x'' \triangle a) \bullet (x''_n \triangle a),\\
&\vdots \qquad\qquad\qquad \vdots
\end{aligned}$$

The sequence of elements $x''_n \triangle a$ is strictly decreasing by (ii) and C5″ and bounded by $x'' \triangle a$. Thus it has a limit. R9 yields $\lim(x_n \triangle a) = (x'' \triangle a) \bullet \lim(x''_n \triangle a)$. By C3″ and C5″ it is inferred that $x'' \triangle a = \lim(x''_n \triangle a)$.

(iv) It follows that there is an index k such that $x'' \triangle a < x''_k \triangle a < b$. C5″ yields $x'' < x''_k$, further $x''_k \in A$. This contradicts the maximality assumption of x'' and therefore proves the existence of an element z with the asserted property.

(v) The uniqueness of z follows from the fact that $z \triangle a = z' \triangle a$ implies $z = z'$ by C5″. Q.E.D.

*R*30. Let u_n and v_n, respectively, denote the general elements of two infinite and strictly monotone sequences of elements of M with limits u and v, respectively. Assume further that the infinite sequence with the general element $u_n \triangle v_n$ is strictly monotone. Then $\lim(u_n \triangle v_n) = (\lim u_n) \triangle (\lim v_n)$.

This theorem can be proved in complete analogy with the proof of R28, $\bullet$ being replaced by $\triangle$. The proof does not involve bisymmetry. Q.E.D.

*C*4″. Consider four elements, x, y, u, v. Then $(x \triangle y) \triangle (u \triangle v) = (x \triangle u) \triangle (y \triangle v)$.

Proof. Set $w = x \triangle u$ and $z = y \triangle v$ and define the following infinite sequences.

$$\begin{array}{ll}
w_1 = w \bullet z, & z_1 = z \bullet w, \\
w_2 = w_1 \bullet z_1, & z_2 = z_1 \bullet w_1, \\
\vdots & \vdots \\
w_{n+1} = w_n \bullet z_n, & z_{n+1} = z_n \bullet w_n, \\
\vdots & \vdots
\end{array}$$

By Def. 6, $\lim w_n = \lim z_n = w \triangle z = (x \triangle u) \triangle (y \triangle v)$. Using R29, these 'equations' may be rewritten as

$$\begin{array}{ll}
w_1 = (x \triangle u) \bullet (y \triangle v) & z_1 = (y \triangle v) \bullet (x \triangle u) \\
\quad = (x \bullet y) \triangle (u \bullet v) & \quad = (y \bullet x) \triangle (v \bullet u) \\
\quad = x_1 \triangle u_1. & \quad = y_1 \triangle v_1. \\
w_2 = (x_1 \triangle u_1) \bullet (y_1 \triangle v_1) & z_2 = (y_1 \triangle v_1) \bullet (x_1 \triangle u_1) \\
\quad = (x_1 \bullet y_1) \triangle (u_1 \triangle v_1) & \quad = (y_1 \bullet x_1) \triangle (v_1 \bullet u_1) \\
\quad = x_2 \triangle u_2. & \quad = y_2 \triangle v_2.
\end{array}$$

Generally, for all n,

$$w_n = x_n \triangle u_n, \qquad z_n = y_n \triangle v_n.$$

In these reformulations the definitions of x_n, u_n, y_n and v_n follow the notation employed in Def. 6 and are therefore not believed to require further explanation. Now it follows from R30 that $\lim w_n = w \triangle z = (x \triangle u) \triangle (y \triangle v) = (\lim x_n) \triangle (\lim u_n) = (x \triangle y) \triangle (u \triangle v)$ as has been asserted. Q.E.D.

The results obtained so far apply when M is of class I (remember Def. 5 and Def. 6). For the alternative of M belonging to class I the following definitions are given:

Def. 6′. If M is of class II then the operation $\vartriangle$ is defined by $x \vartriangle y = \lim x_n$, where

$$\begin{array}{ll} x_1 = y \bullet x, & y_1 = x \bullet y, \\ x_2 = y_1 \bullet x_1, & y_2 = x_1 \bullet y_1, \\ \vdots & \vdots \\ x_{n+1} = y_n \bullet x_n, & y_{n+1} = x_n \bullet y_n, \\ \vdots & \vdots \end{array}$$

Following the patterns of the previous derivations it may be readily shown that analogous results obtain from Def. 6′.

Thus it may be said that in general the operation $\vartriangle$, regardless of the class membership of M, fulfils all assumptions (axioms) C1″ to C6″, therefore exhibiting the same features as did operation $*$. It should also be noted that the definition $\vartriangle$ was based on the introduction of $\bullet$, the latter only lacking the requirement of commutativity.

8. NUMERICAL REPRESENTATIONS

Some of the previous results will now be used in solving the scaling problem for M, i.e. the question of existence and uniqueness of an assignment of numerical values to the elements of M such that an isometric map of M obtains. However, attention will be restricted to *linear* representations because of their numerical convenience.

Def. 7. M will be said to be linearly scalable with respect to order and a concatenation operation $\circ$ if, for every $z \in M$, a real number $s(z)$ can be found such that $a \gtreqless b$ is equivalent to $s(a) \gtreqless s(b)$, respectively, and $c = a \circ b$ is equivalent to $s(c) = \alpha \cdot s(a) + \beta \cdot s(b) + \gamma$, where α, β and γ are real constants.

It has been shown in a previous article (Kristof, 1967 *b*; also Pfanzagl, 1959 *b*) that a linear scale of M can be obtained when the operation $*$ is involved. The expression for $s(x*y)$ was required to be

$$s(x*y) = \tfrac{1}{2}\{s(x) + s(y)\} \tag{1}$$

in order to correspond with one's notion of taking the 'middle' between two elements x and y. It has also been shown that any such scaling function $s(x)$ is necessarily continuous and that the group of admissible transformations, preserving the form of (1), is the group of all positive linear transformations.

It is obvious, of course, that eqn. (1) constitutes the only possible linear expression for $s(x*y)$ in terms of $s(x)$ and $s(y)$. Namely, setting $s(x*y) = \alpha \cdot s(x) + \beta \cdot s(y) + \gamma$ leads to $(\alpha + \beta - 1)s(x) + \gamma = 0$ by means of reflexivity. Due to the arbitrariness of $s(x)$, the left side must vanish identically, which gives $\alpha + \beta = 1$ and $\gamma = 0$. Commutativity then requires $(\alpha - \beta)\{s(x) - s(y)\} = 0$, thus $\alpha = \beta$ since in general $s(x) \neq s(y)$. Therefore $\alpha = \beta = \frac{1}{2}$, which establishes the form of eqn. (1).

This result applies equally when the operation $\vartriangle$ is involved since $\vartriangle$ satisfies all assumptions C1″ to C6″ and therefore is formally indistinguishable from an operation $*$. Hence there is an order-preserving, continuous and essentially unique scale $s(x)$ such that

$$s(x \vartriangle y) = \tfrac{1}{2}\{s(x) + s(y)\}. \tag{2}$$

This is the only possible linear form. This result is used to establish a scale when the operation $\bullet$ is involved. It only remains to secure an expression for $s(x \bullet y)$ in terms of $s(x)$ and $s(y)$ on the basis of the above results. Such an expression *must* exist, since $s(x \bullet y)$ is wholly determined by $s(x)$ and $s(y)$. To this end, R29 is invoked.

$$(x \vartriangle y) \bullet (u \vartriangle v) = (x \bullet u) \vartriangle (y \bullet v). \tag{3}$$

The continuity and monotonicity of $s(z)$ ensures the existence, continuity and monotonicity of the inverse function $s^{-1}(.)$. Applying eqn. (2), the left side of eqn. (3) can be written as

$$s^{-1}\left[\frac{s(x)+s(y)}{2}\right] \bullet s^{-1}\left[\frac{s(u)+s(v)}{2}\right].$$

Similarly, the right side of eqn. (3) becomes

$$s^{-1}\left[\frac{s[s^{-1}\{s(x)\} \bullet s^{-1}\{s(u)\}] + s[s^{-1}\{s(y)\} \bullet s^{-1}\{s(v)\}]}{2}\right].$$

These somewhat unwieldy expressions may be simplified by setting $s(x)=\xi$, $s(y)=\eta$, $s(u)=\mu$, $s(v)=\nu$. Further, both sides are taken as arguments of the function $s(.)$. This gives

$$s\left[s^{-1}\left(\frac{\xi+\eta}{2}\right) \bullet s^{-1}\left(\frac{\mu+\nu}{2}\right)\right] = \frac{s\{s^{-1}(\xi) \bullet s^{-1}(\mu)\} + s\{s^{-1}(\eta) \bullet s^{-1}(\nu)\}}{2}.$$

Now define $S(\xi, \eta) = s\{s^{-1}(\xi) \bullet s^{-1}(\eta)\}$. This function is continuous due to the continuity of $s(.)$, $s^{-1}(.)$, the operation $\bullet$ (C6″) and M. Thus

$$S\left(\frac{\xi+\eta}{2}, \frac{\mu+\nu}{2}\right) = \frac{S(\xi, \mu) + S(\eta, \nu)}{2}. \tag{4}$$

This is Jensen's functional equation for two variables (Aczél, 1961, p. 199). The only continuous solutions are of the form

$$S(\xi, \eta) = \alpha\xi + \beta\eta + \gamma \tag{5}$$

with real constants α, β, γ. Considering the previous substitutions, it follows that

$$s(x \bullet y) = \alpha \cdot s(x) + \beta \cdot s(y) + \gamma.$$

Monotonicity implies $\alpha > 0$ and $\beta > 0$. Reflexivity obviously requires $\alpha + \beta = 1$ and $\gamma = 0$ with $\alpha = 0$ and $\alpha = 1$ excluded. With these provisions, the following linear expression is obtained

$$s(x \bullet y) = \alpha \cdot s(x) + \beta \cdot s(y). \tag{6}$$

Thus $s(x \bullet y)$ is the quasilinear mean of $s(x)$ and $s(y)$. The foregoing derivation followed a pattern set by Aczél (1961, p. 199) in his study of the bisymmetry equation.

The numerical value of α in eqn. (6) is determined by any triple x, y and $x \bullet y$ when $x \neq y$. Namely, eqn. (6) yields

$$\alpha = \frac{s(x \bullet y) - s(y)}{s(x) - s(y)}. \tag{7}$$

It is immediately seen that eqns. 2) and (6) satisfy eqn. (3). Thus eqn. (6) is the desired expression indeed, and it is the only solution based on eqns. (2) and (3) except for positive linear transformations of $s(x)$. However, such transformations evidently do not affect α, which follows immediately from the form of eqn. (7). Thus a linear scale of M *exists* with respect to the operation $\bullet$.

Now it is required to show that all scales linear with respect to the operation $\bullet$ are given by the manifold of scales $s(x)$. To this end, let $t(x)$ be such a scale. Then this must be a monotone and continuous real function of x. Linearity means, in analogy to similar considerations above,

$$t(x \bullet y) = \alpha \cdot t(x) + \beta \cdot t(y) \tag{8}$$

with $\alpha + \beta = 1$, $\alpha \neq 0, 1$, $\alpha > 0$, $\beta > 0$, α a certain real constant. The inverse function $t^{-1}(.)$ exists and it is monotone and continuous.

Using eqn. (3),

$$t^{-1}\{\alpha t(x \vartriangle y) + \beta t(u \vartriangle v)\} = t^{-1}\{\alpha t(x) + \beta t(u)\} \vartriangle t^{-1}\{\alpha t(y) + \beta t(v)\}.$$

Putting $t(x) = \xi$, $t(y) = \eta$, $t(u) = \mu$, $t(v) = \nu$ and introducing the continuous function $T(\xi, \eta) = t\{t^{-1}(\xi) \vartriangle t^{-1}(\eta)\} = t(x \vartriangle y)$, leads to the result

$$\alpha \cdot T(\xi, \eta) + \beta \cdot T(\mu, \nu) = T(\alpha\xi + \beta\mu, \alpha\eta + \beta\nu). \tag{9}$$

It is required to solve this functional equation in two variables. Put $\eta = \nu$, so that $\alpha\eta + \beta\nu = \eta$. Keeping η constant and then setting $f(\xi) = T(\xi, \eta)$, gives

$$\alpha f(\xi) + \beta f(\mu) = f(\alpha\xi + \beta\mu). \tag{10}$$

According to Aczél (1961, p. 69), the general continuous solution of this latter functional equation with $\alpha + \beta = 1$ is

$$f(\xi) = A\xi + D \tag{11}$$

with real constants A and D.

However, when η is not held fixed, then A and D may be functions of η, $A(\eta)$ and $D(\eta)$, thus

$$T(\xi, \eta) = A(\eta) \cdot \xi + D(\eta). \tag{12}$$

Using eqn. (12), eqn. (9) can be brought into the form

$$\alpha D(\eta) + \beta D(\nu) - D(\alpha\eta + \beta\nu) = a\xi\{A(\alpha\eta + \beta\nu) - A(\eta)\} + \beta\mu\{A(\alpha\eta + \beta\nu) - A(\nu)\},$$

where ξ, η, μ and ν are allowed to vary independently. Since the left side of this equation is evidently independent of ξ and μ, so also is the right side. Therefore

$$A(\alpha\eta + \beta\nu) - A(\eta) = A(\alpha\eta + \beta\nu) - A(\nu) = 0, \tag{13}$$

and consequently

$$aD(\eta) + \beta D(\nu) = D(\alpha\eta + \beta\nu). \tag{14}$$

It follows from eqn. (13) that $A(.)$ is simply a constant, A. Then, remembering eqn. (11), $D(\eta)$ must be a continuous function and, again by Aczél (1961, p. 69),

$$D(\eta)=B\eta+C \tag{15}$$

with real constants B and C. So finally, from eqn. (12),

$$T(\xi,\eta)=A\xi+B\eta+C. \tag{16}$$

Considering the previous substitutions, this becomes

$$t(x \vartriangle y)=At(x)+Bt(y)+C. \tag{17}$$

It is immediately found that eqn. (16) solves eqn. (9).

Reflexivity and commutativity of the operation $\vartriangle$ require $A=B=\frac{1}{2}$ and $C=0$. Thus

$$t(x \vartriangle y)=\tfrac{1}{2}\{t(x)+t(y)\}. \tag{18}$$

But it has already been remarked earlier that all scales $t(x)$ or $s(x)$, whatever the notation may be, which obey a formula of the form of eqn. (18) involving the operation $\vartriangle$ are connected by positive linear transformations. Hence $t(x)$ as considered above belongs to the manifold of functions $s(x)$ introduced in eqn. (2). This establishes the *uniqueness* of linear scales involving $\bullet$ up to positive linear transformations. The functions $t(.)$ as used in eqn. (18) are identical with the functions $s(.)$ considered in eqn. (2). Therefore $s(.)$ and $t(.)$ may be used interchangeably.

The class membership of sets M as introduced in Def. 5 expresses itself in the coefficient α in eqn. (8). It is easily seen that $\alpha>\frac{1}{2}$ if M is of class I and $\alpha<\frac{1}{2}$ if M is of class II. At any rate, α is a *fixed constant* and not subject to any transformations as is $t(x)$, i.e. α does not depend upon the choice of a particular $t(x)$.

Next, attention is given to the operation $\circ$. The existence and uniqueness of scales M linear with respect to this operation will be examined. As to the *existence*, an expression will be obtained for $t(x\circ y)$ in terms of $t(x)$ and $t(y)$. Such an expression *must* exist since $t(x\circ y)$ is wholly determined by $t(x)$ and $t(y)$. To this end, recalling Def. 4, write

$$x\circ y=(x\bullet y)\circ(x\bullet y). \tag{19}$$

From eqn. (8),

$$t(x\circ y)=t[t^{-1}\{\alpha t(x)+\beta t(y)\}\circ t^{-1}\{\alpha t(x)+\beta t(y)\}].$$

Hence $t(x\circ y)$ is a function solely of $\alpha t(x)+\beta t(y)$, which is written simply as

$$t(x\circ y)=g\{\alpha t(x)+\beta t(y)\}, \tag{20}$$

with the unknown function $g(.)$. This function is monotone and continuous because $t(.)$ and the operation $\circ$ are monotone and continuous (C6) and M is continuous. Thus the inverse function $g^{-1}(.)$ exists and is monotone and continuous. In order to determine $g(.)$ a special case of C4 is utilized:

$$(x\circ x)\circ(y\circ y)=(x\circ y)\circ(x\circ y). \tag{21}$$

Taking both sides as arguments of the functions $t(.)$ and $g^{-1}(.)$ and using eqn. (20) finally gives

$$\alpha g\{t(x)\}+\beta g\{t(y)\}=g\{\alpha t(x)+\beta t(y)\}. \tag{22}$$

This functional equation has the same form as eqn. (10), hence its general continuous solution is (Aczél, 1961, p. 69)

$$g\{t(x)\}=Et(x)+H,$$

E and H being real constants. Using eqn. (20), it follows that

$$t(x \circ y)=Ft(x)+Gt(y)+H, \tag{23}$$

where $F=E\alpha$ and $G=E\beta$. However, eqns. (23) and (8) must together satisfy $t(x \circ y)=t\{(x \bullet y) \circ (x \bullet y)\}$ according to eqn. (19). This leads to

$$(F\alpha+G\alpha-F)t(x)+(F\beta+G\beta-G)t(y)=0,$$

thus $F\alpha+G\alpha-F=F\beta+G\beta-G=0$ because of the arbitariness of $t(x)$ and $t(y)$. This leads to the constraint

$$\frac{F}{G}=\frac{\alpha}{\beta}=\frac{\alpha}{1-\alpha}. \tag{24}$$

Another constraint is imposed by eqn. (23) itself when any triple x, y and $x \circ y$ with $x \neq y$ is chosen so that $t(x)$, $t(y)$ and $t(x \circ y)$ are given values. Finally, there is a third constraint due to the existence of precisely one reflexive element, z say, as required by R11 (no consideration is given here to the case in which all of the elements of M are reflexive because this has been covered by eqn. (8) involving the operation $\bullet$). All three constraints are formulated as the equations:

$$\left.\begin{aligned} F(1-\alpha)-G\alpha&=0,\\ Ft(x)+Gt(y)+H&=t(x \circ y),\\ Ft(z)+Gt(z)+H&=t(z), \end{aligned}\right\} \tag{25}$$

arranging for $x \circ y \neq z$, which is of course possible to achieve. This system of inhomogeneous linear equations with the unknowns F, G, H possesses *precisely one solution* because the determinant of the matrix of coefficients does not vanish:

$$\begin{vmatrix} 1-\alpha & \alpha & 0\\ t(x) & t(y) & 1\\ t(z) & t(z) & 1 \end{vmatrix}=\alpha t(x)+(1-\alpha)t(y)-t(z)=t(x \bullet y)-t(z).$$

But $x \circ y \neq z$ implies $t(x \bullet y)-t(z) \neq 0$. For, suppose that $x \bullet y=z$. Then, recalling eqn. (19), $x \circ y=(x \bullet y) \circ (x \bullet y)=z \circ z=z$, i.e. $x \circ y=z$. This contradiction proves that $x \bullet y \neq z$ and, by monotonicity of $t(.)$, $t(x \bullet y)-t(z) \neq 0$. Thus eqn. (25) possesses precisely one solution F, G, H indeed. In other words, these coefficients are wholly determined by the choice of a particular $t(.)$.

Hence the *existence* of a scale of M linear with respect to the operation $\circ$ has been established.

Now consideration is given to the *uniqueness* question. In particular, it is required to show that all scales linear with respect to the operation $\circ$ are given by the manifold of scales $t(x)$ and so, equivalently, by the manifold of scales $s(x)$. To this end, let $r(x)$ be such a scale. Then this must be a monotone and continuous real function of x. Linearity means

$$r(x \circ y) = Fr(x) + Gr(y) + H \tag{26}$$

with some real constants F, G, H. These constants must obey a certain condition obtained from eqn. (19). Namely, $r(x \circ y) = r\{(x \bullet y) \circ (x \bullet y)\}$, which becomes, upon applying eqn. (26),

$$r(x \bullet y) = \frac{F}{F+G} r(x) + \frac{G}{F+G} r(y) . \tag{27}$$

Setting $\alpha = F/(F+G)$ and $\beta = G/(F+G)$ gives $a + \beta = 1$ and, of course, $\alpha \neq 0, 1$, $\alpha > 0$, $\beta > 0$. Therefore $r(x)$ is a particular case contained in the manifold of scales $t(x)$ as introduced in eqn. (8). Hence the manifold of scales $r(x)$ is identical with the manifold of scales $t(x)$ and thus with the manifold of scales $s(x)$. This manifold was given by all positive linear transformations of a particular $s(x)$. This result answers the uniqueness question.

It has already been said that α in eqn. (8) is a fixed constant independent of the particular choice of $t(x)$ and thus signifies a structural aspect of the set M. Now it is necessary to examine in which way F, G and H in eqn. (23) are affected if $r(.)$ is subjected to positive linear transformations (in Def. 7 the symbols α, β, γ were chosen). Consequently, eqn. (26) is considered along with

$$r'(x \circ y) = \mathrm{F}'r'(x) + G'r'(y) + H'$$

when $r'(.) = \lambda r(.) + \mu$, $\lambda > 0$, λ and μ real constants. It is easily found that

$$r(x \circ y) = F'r(x) + G'r(y) + \frac{(F' + G' - 1)\mu + H'}{\lambda} .$$

Equating this with eqn. (26) gives

$$F' = F; \quad G' = G; \quad H' = \lambda H - (F + G - 1)\mu. \tag{28}$$

Hence F and G are *not affected* by admissible transformations of $r(.)$. Evidently, by a suitable choice of λ and μ, H' will assume any prescribed value. In particular, it can be made zero.

This paper establishes a formal theory. Special applications will be considered in another article. For example, the derivation of scales for psychophysical continua may involve an operation $\bullet$ when a time-order error enters into subjective comparisons of stimuli. If, in addition, there is a tendency to avoid extreme responses in experiments which require the selection of stimuli 'halving' perceptual intervals, then, perhaps, an operation $\circ$ will sufficiently well describe the situation. It has been shown that such 'irregularities' do not necessarily preclude scalability of a domain.

REFERENCES

ACZÉL, F. (1961). *Vorlesungen über Funktionalgleichungen und ihre Anwendungen.* Stuttgart and Basel: Birkhäuser Verlag.

KRISTOF, W. (1967 *a*). A foundation of interval scale measurement. *Research Bulletin,* 67–22. Princeton, N.J.: Educational Testing Service.

KRISTOF, W. (1967 *b*). Measurement based on halvability of intervals. *Research Bulletin,* 67–18. Princeton, N.J.: Educational Testing Service.

LUCE, R. D. & TUKEY, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *J. math. Psychol.* **1,** 1–27.

PFANZAGL, J. (1959 *a*). A general theory of measurement—applications to utility. *Naval Res. Logist. Q.* **6,** 282–294. (Also in R. D. Luce, R. R. Bush & E. Galanter (eds.), *Readings in Mathematical Psychology*, vol. 2. New York: Wiley, 1965.)

PFANZAGL, J. (1959 *b*). *Die axiomatischen Grundlagen einer allgemeinen Theorie des Messens.* Würzburg: Physica-Verlag.

TVERSKY, A. (1967). A general theory of polynomial conjoint measurement. *J. math. Psychol.* **4,** 1–20.

Vol. 21
Part 2
231–238

The British Journal of Mathematical and Statistical Psychology

November
1968

AN EVALUATION OF RASCH'S STRUCTURAL MODEL FOR TEST ITEMS

By J. ANDERSON
University of New England
G. E. KEARNEY
Papua and New Guinea Psychological Research Unit
and A. V. EVERETT
University of Queensland

Rasch's item analysis model for intelligence tests was examined using a 45-item spiral omnibus intelligence test, administered to two samples of 608 and 874 subjects respectively. Five separate hypotheses were tested. Findings provide evidence for Rasch's claim that the difficulty level of items and ability level indices are independent of the sample on which they are based.

1. INTRODUCTION

Rasch (1960) proposed three probabilistic models for use in test construction. The first is concerned with accuracy in reading tests or, as Rasch calls it, a 'model for misreading'; the second is designed for speeded tests, or more specifically for reading speeds; and the third for item analysis of intelligence test items. It is this third of Rasch's models with which this paper is concerned.

The noteworthy feature of Rasch's probabilistic models is that he abolishes the restriction which limits any findings to a particular sample or population. Individual ability estimates are derived which, Rasch claims, are independent of the difficulty of particular test items or of a particular test. Item difficulty estimates are also derived, which are likewise claimed to be independent of the particular group of subjects involved. No recourse is made to the normal distribution or to any other distribution. Rasch concludes that 'the concepts of measurement introduced through the definition of the two types of parameters differ radically from those employed in the psychometric theory of mental tests'. Rasch's work attempts 'a new approach to test-psychology'.

To date, Rasch's work has attracted little attention. Besides Rasch's own writings (1960, 1966) the authors have sighted only one unpublished paper and one review of Rasch's work. Before examining, then, how well Rasch's model fits empirical data, it seems necessary first to describe briefly the model and Rasch's control of the model.

The model is applicable to any situation where a group of individuals respond to each of a number of items, responses being marked either right or wrong.

Rasch (1966) has recently presented the model in more generalized form, but the concern here is with zero-one data. Two parameters are postulated: a_i, ability level of individual i or of group of individuals of the same ability i; and d_j, difficulty level of item j. Both parameters are assumed to be monotonic increasing functions of ability and of difficulty respectively.

Rasch makes two assumptions. Firstly, that given the values of the parameters, all answers are stochastically independent. Secondly, that the probability of individual i (or group i) answering item j correctly is a function of the ratio a_i/d_j, i.e. between the degree of ability of the individual and the degree of difficulty of the item; but the probability is not dependent on the values of the parameters a and d separately. The necessary properties of the function are that it has the range 0 to 1 as a_i/d_j, theoretically at least, has the domain 0 to ∞. Quite arbitrarily, Rasch chose the function $x/(1+x)$. Hence, the probability of individual i correctly answering item j is given by:

$$P_{ij} = \frac{a_i/d_j}{1+a_i/d_j} = \frac{a_i}{a_i+d_j}. \tag{1}$$

Similarly, the probability of individual i not answering item j correctly is:

$$1 - P_{ij} = d_j/(a_i+d_j). \tag{2}$$

From eqns. (1) and (2) Rasch derived his basic model:

$$P_{ij}/(1-P_{ij}) = a_i/d_j. \tag{3}$$

To check the validity of this model Rasch suggests a relatively simple control which, however, he admits rests upon approximations, about the exact nature of which little is known. Again two assumptions are made. Firstly, it is assumed that the best estimate of the ability parameter a for an individual (or for a group of individuals) is his raw score. Secondly, it is assumed that the best estimate of the item difficulty parameter d is the frequency of correct responses by individuals doing the test. Keats (1967) explains the model in terms of a matrix where ordered groups of subjects define the rows and items define the columns, and the entries are the proportions P_{ij} of subjects in group i answering item j correctly. For such a matrix it is possible to estimate the probabilities in eqns. (1) and (2) and from eqn. (3) thus obtain:

$$P_{ij}/(1-P_{ij}) \approx a_i/d_j. \tag{4}$$

The approximation sign is used because of errors of estimation (e.g. the misclassification of subjects into ability levels). Except for random variations, eqn. (4) should hold for all $1 > P_{ij} > 0$. Where $P_{ij} = 0$ or 1, it is necessary either to delete columns and/or rows or to group subjects and/or items, but eqn. (4) is still assumed valid within some approximation.

Taking logarithms of both sides of eqn. (4) and denoting this logistic transformation by l_{ij} we obtain:

$$l_{ij} \approx \log\{P_{ij}/(1-P_{ij})\} \approx \log a_i - \log d_j. \tag{5}$$

The logistic transform matrix is shown in Table 1. For all m items in Table 1 the difference between pairs of subject groupings is the same. For example, for the first two subjects:

$$l_{1j} - l_{2j} = \log a_1 - \log d_j - (\log a_2 - \log d_j) = \log a_1 - \log a_2, \tag{6}$$

and the differences between the left-hand sides of eqn. (5) should likewise be approximately equal to the difference in eqn. (6). Graphically, this means that if we plot row 1 against row 2, points will lie around a straight line of unit slope and the intercept on the ordinate axis yields an estimate of eqn. (6). Because each ability group is based on relatively few observations, the precision of these estimates obtained from comparisons of pairs of the n groups can be increased by plotting each ability group against the mean of all ability groups. The degree to which data tend to fall on straight lines of unit slope is in effect the test of the goodness of fit of the Rasch model and, if achieved, column means provide estimates of the difficulty of items plus an arbitrary constant while row means provide estimates of ability groupings plus an arbitrary constant.

TABLE 1. LOGISTIC TRANSFORMS OF PROPORTIONS OF ORDERED GROUPS OF SUBJECTS GIVING THE 'CORRECT' RESPONSES TO ITEMS

				Items			Means
		1	…	j	…	m	
	1	$\log a_1 - \log d_1$	…	$\log a_1 - \log d_j$	…	$\log a_1 - \log d_m$	$\log a_1 - \log d$
	2	$\log a_2 - \log d_1$	…	$\log a_2 - \log d_j$	…	$\log a_2 - \log d_m$	$\log a_2 - \log d$
	⋮	⋮		⋮		⋮	⋮
Subjects	i	$\log a_i - \log d_1$	…	$\log a_i - \log d_j$	…	$\log a_i - \log d_m$	$\log a_i - \log d$
	⋮	⋮		⋮		⋮	⋮
	n	$\log a_n - \log d_1$	…	$\log a_n - \log d_j$	…	$\log a_n - \log d_m$	$\log a_n - \log d$
Means		$\log \bar{a} - \log d_1$	…	$\log \bar{a} - \log d_j$	…	$\log \bar{a} - \log d_j$	

Brooks (1965), using two different ability levels (school grades), tested the model by means of both verbal and non-verbal items from the Lorge–Thorndike Intelligence Test and found that 27·2 per cent of items checked failed to fit the model at the 0·05 level. Brooks, however, did not use items that were completely common to both ability groups.

The model has also been tested by Kearney (1966), who found that 20·4 per cent of items tested failed to fit the model at the 0·05 level (4·08 per cent at the 0·01 level).

In neither study were the indices recalculated after items that had failed to meet the model were removed, a procedure suggested by Brooks at the conclusion of his paper.

In the present study the Rasch model was examined to investigate the following hypotheses: (1) Item difficulty indices are independent of the sample on which they are based. (2) Indices of item difficulty are not substantially influenced by other items in the test. (3) Item difficulty indices are more stable when only items that fit the model are considered. (4) Indices of item difficulty are not influenced by the composition of ability groupings. (5) Ability level indices are independent of the sample on which they are based.

2. METHOD

The test used was a 45-item spiral omnibus intelligence type test that is used for screening recruits who apply to join either the Australian Army or the Royal Australian Navy. It is used for all applicants except those who apply for officer training courses.

The test results of two samples were investigated. The first sample consisted of 608 recruit applicants to the Citizen Military Forces (CMF), which is a part-time system of military training; the other, 874 recruit applicants to the Royal Australian Navy (RAN). The second sample consisted of 446 General Service recruits, 129 Reservists (the Navy equivalent of the CMF) and 279 recruits to the women's section of the RAN.

The test is known to have high internal reliability. Tucker (1966) found a reliability coefficient of 0·916, using the Spearman–Brown split-half method on a sample of 507 recruit applicants to the Regular Army. The reliability was checked for each sample using Kuder–Richardson's formula 20. The coefficients obtained were 0·921 for the CMF sample, 0·913 for the RAN sample, and 0·910, 0·908 and 0·902 for the three RAN subsamples respectively.

The initial procedure was to form a frequency matrix of the number of times an item was correct at each raw score level. Raw score levels were then grouped into six ability levels, corresponding to the selection groupings used by the Army to determine cut-off points. Because the Rasch model does not hold for $P_{ij}=1$ and $P_{ij}=0$, the matrix was further reduced by deleting those items (rows) in which all subjects of a given ability level passed the item. Similarly, in those cases where no subject of a given ability level passed the item, the row or item was deleted. In all, 12 test items were not examined and on *a priori* grounds did not fit the model.

This reduced 6×33 matrix was converted to form a proportion matrix. The logarithms of $P/(1-P)$ were then calculated. The logistic transforms (l_{ij}) for the CMF sample are shown in Table 2.

Item scale values and ability group scale values were calculated in the manner described in the previous section, i.e. from row and column means a constant was subtracted (this is just a question of choice of unit) and antilogarithms taken. These results for the CMF sample appear in Table 2.

Rasch tested the goodness of fit of his model by visual examination of the graphs of items (i.e. the columns of Table 1) against the means of all items (i.e. the row means of Table 1). Here, following Brooks (1965), t tests were calculated to investigate whether the regression of the six logistic transformations for each item departed significantly from unit slope. It had been hoped to treat each of the RAN subsamples separately, but too many cell entries contained proportions of zero or unity to make the analysis worth while. Consequently

TABLE 2. LOGISTIC TRANSFORMS, ITEM AND ABILITY SCALE VALUES FOR CMF SAMPLE

Items	Ability groups 1	2	3	4	5	6	Row means	Item scale values
5	0·699	1·355	1·672	2·076	1·663	1·813	1·541	0·0546
6	0·184	0·785	0·805	0·915	1·405	1·505	0·933	0·2213
7	0·527	1·355	0·934	1·362	1·505	1·633	1·220	0·1146
8	0·477	0·785	1·260	1·362	1·405	1·633	1·154	0·1330
10	0·184	0·785	0·845	1·146	1·322	2·117	1·067	0·1629
11	−0·073	0·384	0·527	0·996	1·322	1·633	0·792	0·3062
12	−0·261	−0·186	0·073	0·269	0·570	0·926	0·232	1·1120
13	−0·261	−0·186	0·407	0·580	0·926	1·405	0·478	0·6310
14	−0·184	0·212	0·637	1·041	1·086	1·813	0·768	0·3236
15	−0·527	0·061	0·342	0·726	0·860	1·505	0·495	0·6067
16	−0·342	0·212	0·364	0·845	1·086	1·322	0·581	0·4977
17	−0·301	0·377	0·340	1·146	1·190	1·813	0·776	0·3177
18	−0·527	0·111	0·128	0·649	0·962	1·190	0·419	0·7228
19	−0·342	−0·111	0·364	0·649	0·748	1·322	0·439	0·6918
20	−0·527	−0·212	0·281	0·726	0·962	1·633	0·477	0·6324
21	−0·580	−0·377	0·241	0·497	0·892	1·190	0·311	0·9268
22	−1·362	−0·037	0·073	0·439	0·774	1·505	0·232	1·1120
23	−1·362	−0·536	−0·241	0·206	0·590	1·505	0·027	1·7820
24	−0·934	−0·785	−0·128	0·368	1·136	1·633	0·215	1·1560
25	−0·580	−0·348	−0·109	0·517	0·892	1·505	0·313	0·9226
26	−0·768	−0·838	−0·203	0·285	0·723	1·086	0·048	1·6980
27	−1·176	−0·896	−0·453	−0·014	0·393	0·774	−0·229	3·2140
28	−0·845	−0·785	−0·241	0·334	1·136	2·117	0·286	0·9954
29	−0·934	−0·896	−0·453	0·237	0·860	1·633	0·075	1·6000
30	−1·176	−1·224	−0·453	0·131	0·723	1·000	−0·167	2·7860
31	−1·672	−0·838	−0·036	0·403	1·086	2·117	0·177	1·2620
32	−1·176	−0·961	−0·385	−0·041	0·286	0·723	−0·255	3·4120
33	−1·041	−0·785	−0·430	−0·269	0·053	0·426	−0·341	4·1590
34	−1·672	−1·845	−0·934	−0·915	−0·460	0·160	−0·945	16·7100
35	−1·176	−0·961	−0·241	0·301	0·570	1·405	−0·017	1·9720
36	−1·041	−0·896	−0·580	−0·301	−0·013	0·802	−0·338	4·1300
37	−1·672	−1·224	−1·260	−0·580	−0·286	0·272	−0·792	11·7500
38	−1·672	−0·896	−0·699	−0·673	−0·620	−0·291	−0·792	11·7500
Column means:	−0·670	−0·286	0·077	0·468	0·780	1·301		
Ability scale values:	0·113	0·273	0·630	1·549	3·117	10·520		

the three RAN subsamples were treated as one unit. Scale values for 33 items for the CMF and for the RAN samples, together with the significance of the departure from unit slope, are presented in Table 3.

To test the first hypothesis, scale values derived from the CMF and RAN samples were correlated, using the product-moment statistic. A coefficient of

$r = 0.958$ was obtained, evidence for Rasch's claim that scale values for items are independent of the sample.

To test the second hypothesis, the items that departed from unit slope were removed. This was done separately for those items significant at $P < 0.01$ and those significant at $P < 0.05$. It is to be noted here that the probability of

TABLE 3. (*a*) ITEM SCALE VALUES AND SIGNIFICANCE OF DEPARTURE FROM UNIT SLOPE. (*b*) ITEM SCALE VALUES RECALCULATED WITH ITEMS FAILING TO FIT THE MODEL REMOVED

	CMF				RAN			
Item	Scale values of all items	*t*	Scale values		Scale values of all items	*t*	Scale values	
			<0·01	<0·05			<0·01	<0·05
5	0·06	2·32	0·06	0·05	0·04	4·25*	0·04	—
6	0·22	3·50*	0·23	—	0·20	7·38**	—	—
7	0·12	3·26*	0·12	—	0·07	0·95	0·07	0·05
8	0·13	4·55*	0·14	—	0·10	3·40*	0·10	—
10	0·16	1·22	0·17	0·14	0·22	6·71**	—	—
11	0·31	2·42	0·32	0·27	0·36	1·72	0·35	0·25
12	1·11	7·52**	—	—	1·11	4·80**	—	—
13	0·63	1·66	0·67	0·55	0·64	2·81*	0·63	—
14	0·32	0·40	0·34	0·28	0·40	3·80*	0·40	—
15	0·61	0·52	0·64	0·53	0·64	2·02	0·64	0·45
16	0·50	2·02	0·53	0·44	0·53	4·65**	—	—
17	0·32	0·24	0·34	0·28	0·37	0·35	0·37	0·26
18	0·72	1·47	0·77	0·63	0·64	0·51	0·64	0·45
19	0·69	3·04*	0·73	—	0·76	2·57	0·75	0·53
20	0·63	2·75	0·67	0·56	0·81	0·47	0·81	0·57
21	0·93	0·61	0·98	0·81	1·24	0·18	1·22	0·86
22	1·11	1·46	1·18	0·98	1·35	1·04	1·33	0·94
23	1·78	4·08*	1·89	—	1·72	2·59	1·70	1·20
24	1·16	3·57*	1·23	—	0·89	3·27*	0·87	—
25	0·92	1·23	0·98	0·81	1·02	3·43*	1·01	—
26	1·70	0·58	1·80	1·49	1·55	2·70	1·53	1·08
27	3·21	0·89	3·40	2·81	3·16	1·27	3·12	2·20
28	1·00	3·27*	1·07	—	0·85	7·83**	—	—
29	1·60	2·70	1·69	1·40	1·83	13·34**	—	—
30	2·79	1·82	2·95	2·44	3·08	4·65**	—	—
31	1·26	12·30**	—	—	0·78	6·64**	—	—
32	3·41	0·17	3·61	2·99	2·89	4·60*	2·85	—
33	4·16	8·44**	—	—	6·53	0·99	6·44	4·55
34	16·71	0·04	17·70	14·62	18·66	0·53	18·41	13·00
35	1·97	4·40*	2·09	—	1·65	4·42*	1·63	—
36	4·13	0·84	4·39	3·62	5·05	0·21	4·98	3·52
37	11·75	0·15	12·45	10·28	16·67	1·31	16·44	11·64
38	11·75	2·70	12·47	10·30	7·60	7·88**	—	—

* $P < 0.05$; ** $P < 0.01$.

being significant at the 0·05 level is the more stringent test, as it is the failure to fit the model that is being tested. With these items removed, the analysis was repeated and two new sets of scale values were obtained for each sample. These recalculated scale values are shown in Table 3. The correlation between the original scale values (CMF and RAN) and the two new scale values was $r=0{\cdot}9999$ for all four cases.

This seems to indicate that the recalculation of scale values after those items not fitting the model have been removed is unnecessary.

The third hypothesis was examined by correlating the scale values for those items that fit the model in both samples. The coefficient obtained was $r=0{\cdot}990$. The difference between this coefficient and that obtained when all items were correlated ($r=0{\cdot}958$) is 0·032, which is in the expected direction although not significant.

The fourth hypothesis was examined for only the CMF sample. Six different ability groupings were formed and the resultant scale values were correlated with those obtained for the ability groupings in the original CMF analysis. The correlation between the item weights was $r=0{\cdot}996$.

The fifth hypothesis was examined by correlating the six ability scale values obtained for the CMF and RAN samples. The coefficient obtained was $r=0{\cdot}992$.

3. DISCUSSION

Invariance in Rasch's model appears to be established, in that neither the sample from which the scale values were derived, nor the presence of items that failed to meet the model, appears to have any real influence on the resultant scale values.

One problem that arises is that there is some difficulty in evaluating the level at which an item is accepted as fitting the model. As can be seen from Table 3, 11 items are rejected at the 0·05 level in the CMF sample and 17 in the RAN sample. Of these, seven are common. At the 0·01 level, three are rejected from the CMF sample and nine from the RAN sample. Of these, two are common.

There were 17 items rejected at the 0·05 level when the CMF data were analysed, using the different ability groupings. Of these, seven items were rejected at the 0·01 level.

Initially, these discrepancies appear disconcerting. The difference in t values between all items on the CMF sample analysed by both ability groupings is only 0·207. It appears that the difference is small, but is sufficient to make another six items significant at the 0·05 level and four at the 0·01 level. Part of this increase can be caused, as Rasch suggests, by the ability groupings being too coarse.

It is suggested that the 0·05 level of significance used by Brooks (1965) is probably too stringent a measure, and the 0·01 level is more appropriate when testing the fit of the model. If the 0·01 level is used, 10 per cent of items fail to fit the model for the CMF sample and 33·3 per cent for the RAN sample. It is to be noted that the test was not designed to fit Rasch's model and, with this reservation, the fit could be considered as reasonable.

The model does not lend itself to small samples, and the three RAN subsamples were considered too small for application of this technique.

ACKNOWLEDGEMENTS

This research was carried out while the authors were members of the 21st Unit (CMF) of the Australian Army Psychological Corps. The analysis was carried out on a GE 225 computer, using programs written by the authors.

REFERENCES

BROOKS, R. D. (1965). An empirical check on Rasch's model for item difficulty indexes. (State University of New York, Buffalo.)

KEARNEY, G. E. (1966). The cognitive ability of Aboriginal Australians. (Unpublished Ph.D. dissertation, University of Queensland.)

KEATS, J. A. (1967). Test theory. *Ann. Rev. Psychol.* **18**, 217–238.

RASCH, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Education Research.

RASCH, G. (1966). An item analysis which takes individual differences into account. *Br. J. math. statist. Psychol.* **19**, 49–57.

TUCKER, J. D. (1966). A study of Test RO3. (Unpublished Research Report of 21st Psychology Unit.)

Vol. 21
Part 2
239–246

The British Journal of Mathematical and
Statistical Psychology

November
1968

AN EMPIRICAL STUDY OF GENERALIZABILITY COEFFICIENTS FOR UNMATCHED DATA

By David M. Shoemaker
Oklahoma State University
and H. G. Osburn
University of Houston

Operating within the framework of the theory of generalizability, the main objective of the investigation was that of empirically studying the effect of item difficulty and item content stratification on the gamma coefficients using an unmatched item design. The principal results were as follows: (*a*) the reliability of an unstratified test using the unmatched item design was found to be lower than the reliability of a comparable test using the matched item design, and (*b*) stratification on item difficulty was found to be an extremely important factor in the unmatched item design. These and other results obtained on the gamma coefficients were contrasted with previously reported results for the alpha coefficients involving a matched item design.

1. Introduction

Cronbach, Rajaratnam and Gleser have developed in two major articles (Cronbach *et al.*, 1963; Rajaratnam *et al.*, 1965) a theory of generalizability in which random and stratified random sampling of items is presented as a less restrictive alternative to the assumption of equivalence of parallel forms in classical reliability theory. (It should be noted that, while the present study was mainly concerned with the sampling of items, the theory of generalizability also applies to the sampling of observers, occasions, etc.) According to their theory, 'reliability' is synonymous with 'generalizability'. An individual's test score is never regarded as interpretable in its own right; rather, it is regarded as a sample from a universe of test scores that the individual would have received if he had been tested over all possible random samples of k items from an appropriate population. Any conclusion an investigator draws concerning a particular test score will be generalized to the universe of comparable test scores. To determine the dependability of such a generalization, he must determine how closely a particular test score agrees with the result to be expected from testing the individual over all possible samples of k items from the item population—the correlation between test score and true score.

In the theory of generalizability two coefficients, alpha and gamma, are derived as estimates of the correlation between test score and true score. Each of these coefficients has two forms—one appropriate for the stratified item case and one appropriate for the unstratified item case. The derivations of these coefficients are given by Cronbach *et al.* (1963) and Rajaratnam *et al.* (1965). The alpha coefficients are the appropriate statistics when the data are matched, i.e. all people respond to the same set of k test items drawn at random from a population of admissible items; the gamma coefficients are appropriate for the unmatched data case where k test items are randomly selected for each person independently of the items selected for every other person.

In an empirical study of the matched data case, Cronbach *et al.* (1965) have shown that, for tests constructed by stratification of item content, alpha considerably underestimates the true reliability whereas alpha-stratified provides a good estimate of the true reliability of the test. However, for tests constructed by stratification of item difficulty, alpha is almost as good an estimate of the true reliability as alpha-stratified. In other words, stratification produces a large discrepancy between alpha and alpha-stratified when the stratification is of item content, but a small discrepancy when stratification is of item difficulty.

The importance of unmatched data experiments

The technique of collecting unmatched data has rarely been used in experiments involving ability or achievement tests. Nevertheless, the unmatched data experiment is potentially important. Cronbach (1963), for example, in discussing the measurement of proficiency for evaluating improvements in courses has advocated administering different questions to different students. Lord (1962) has published an empirical study showing the advantages of collecting unmatched data in estimating a distribution of norms. More generally, it can be shown that unmatched data should be collected in situations in which one is interested in the average score of a group of individuals over the universe of content. This is because the standard error of the group mean score is smaller when unmatched data are collected (Lord, 1955; Osburn, 1967). The situation can be summarized as follows: insofar as the investigator is interested in the absolute score of an individual, it does not matter whether matched or unmatched data are collected. If he wants relative scores for individuals, then matched data are required. If he wishes to estimate the mean for a group of persons and proposes to generalize over persons and items, then unmatched data would be preferred.

Purpose of study

The present study deals with the unmatched data case using a design very similar to that used by Cronbach *et al.* (1965). The objective is that of studying the effect of stratification on the gamma coefficients using unmatched items as compared with the corresponding alpha coefficients using matched items.

2. THE GAMMA AND GAMMA-STRATIFIED COEFFICIENTS

Person p, responding to item i, earns observed score X_{pi}. The sum of X_{pi} over all k items is his test score X_{pt}. Person p has a universe score M_p, defined as the mean of X_{pi} over all items in the universe; kM_p is the mean of X_{pt} over the family of tests. For the unmatched data case, the needed coefficient of generalizability is $\rho_{M_pX_{pt}}{}^2$ for data from all conditions pooled. Where each person is observed under k different conditions, mean squares for persons and within-persons are obtained by a one-way analysis of variance, and the generalizability of the total score is

$$\hat{\gamma} = (MS_p - MS_{wp})/MS_p. \tag{1}$$

For dichotomous data the formula reduces to an unbiased version of the Kuder–Richardson formula 21:

$$\hat{\gamma} = \frac{k}{k-1}\left[1 - \frac{1}{nk} - \frac{k\bar{p}(1-\bar{p})}{s_t{}^2}\right], \tag{2}$$

where k refers to the number of items, n to the number of people, and $s_t{}^2$ to the variance of the total test scores for persons.

The formula for the gamma-stratified coefficient is as follows:

$$\hat{\gamma}_s = 1 - \frac{\sum^{h} k_h MS_{wp(h)}}{s_t{}^2}. \tag{3}$$

This equation has a special form for dichotomous items

$$\hat{\gamma}_s = 1 - \frac{\sum^{h} \frac{1}{k_h - 1}[n\bar{X}_{\cdot h}(k_h - \bar{X}_{\cdot h}) - (n-1)s_h{}^2]}{ns_t{}^2}, \tag{4}$$

where $\bar{X}_{\cdot h}$ refers to the average score over persons for stratum h, k_h to the number of items in stratum h, $s_h{}^2$ the variance of the stratum h scores, $s_t{}^2$ the variance of all the scores, and n, the number of people taking the test. This is an unbiased, stratified version of KR 21 using the mean, $\bar{X}_{\cdot h}$, and the estimated variance of subtest scores.

3. METHOD

A procedure was developed that simulates the administration of a single test item to a single examinee. The result is a simulation model of great flexibility that allows for the sampling of both items and individuals.

The model assumes a normally distributed standardized latent ability continuum. The probability of an examinee answering an item correctly is a normal-ogive function of his ability level. Denoting the probability of person j passing the ith item by P_{ij}, this relation is more explicitly stated as

$$P_{ij} = \int_{-\infty}^{(C_j - A_i)/B_i} N(y)\,dy, \tag{5}$$

where C_j is a normally distributed measure of ability, A_i and B_i are values characterizing the item, and $N(y)$ is the standardized normal frequency function.

In order to use eqn. (5), the values of A_i and B_i must be determined for each item. Tucker (1946) has given the equations necessary for expressing A_i and B_i in terms of the item difficulty and the biserial correlation between the item and the underlying ability continuum (R_i). His results are as follows:

$$R_i = (1 + B_i^2)^{-1/2}. \tag{6}$$

The item difficulty level, M_i, over the entire population of examinees is given by

$$M_i = \int_{H_i}^{+\infty} N(y)\, dy, \tag{7}$$

where H_i is a normally distributed variable and $H_i = A_i R_i$.

In the present study, R_i^2 was set at a predetermined value; B_i was consequently determined. The item difficulty parameter H_i was randomly selected from a normal distribution of H_i values, and A_i was consequently determined. The distribution of H_i values was assigned zero mean and unit variance. Lawley (1943) has shown that this assures a rectangular distribution of item difficulties (M_i values).

For single ability tests, setting R_i^2 to a particular value for all items makes the inter-item tetrachoric correlations a constant since $r_{ij} = R_i R_j$. Thus, given R_i^2 and H_i, it was possible to determine A_i, B_i, M_i and r_{ij} for a particular item. A test, then, was a composite of k items with fixed R_i^2 whose H_i values had been randomly or stratified randomly selected from a normal distribution of H_i values.

Content stratification was simulated by assuming a general factor plus a specific factor for each stratum. If the test items were to be stratified on content, the examinee's ability level was computed for each stratum using the equation

$$C_{hi} = W_0 F_{0i} + W_h F_{hi},$$

where F_{0i} and F_{hi} are randomly sampled from the standardized normal distribution and W_0 and W_h are selected in such a way that r_w, the within-stratum inter-item correlation, and r_b, the between-strata inter-item correlation, have predetermined values.

The computer simulation model assigns a score of unity or zero on each test item for every individual in the sample. This is done by computing P_{ij}, the probability of the jth individual passing the ith item, and selecting unity with probability P_{ij} or a zero with probability $1 - P_{ij}$.

Using the computer simulation model described above, hypothetical tests were constructed according to the following specifications: (*a*) each test contained 18 items, (*b*) each test had three strata, (*c*) each stratum contained six items, and (*d*) the probability of passing each test item ranged from 0·01 to 0·99—a 'wide range' of item difficulties.

The formation of strata was accomplished by appropriately setting the values of r_w, the average tetrachoric correlation between items within a stratum, and r_b, the average correlation between items across strata. Values of r_w ($r_w = 0{\cdot}30$, 0·50, 0·70 and 0·90), r_b ($r_b = 0{\cdot}00$, 0·30 and 0·50) and sample size ($N = 25$, 75, 100, 150 and 200) were manipulated for tests stratified on content and on content and difficulty. When tests were constructed with items stratified only according to difficulty level, $r_b = r_w$; and only values of r_w and N were systematically varied. Because of the large amount of computer time involved in generating unmatched data, there were only five replications of each combination of stratification pattern, r_w, r_b and N. The necessary values of ρ_{MX}^2 were computed by numerical integration based on equations given by Lawley (1943).

4. RESULTS AND DISCUSSION

The present study focused on three general questions: (1) How does $E(\rho_{Mt}^2)$, the reliability of tests employing the matched item design, compare with ρ_{MX}^2, the reliability of tests using unmatched items? (2) What is the effect of stratification in unmatched item studies? (3) What effect does stratification have on the sampling variability of gamma over samples of examinees?

TABLE 1. COMPARISON OF MATCHED VERSUS UNMATCHED TEST EXPERIMENTS: RANDOMLY SAMPLED EIGHTEEN-ITEM TESTS

r_w	ρ_{Mt}^2 (From Cronbach *et al.*, 1965)	ρ_{MX}^2
1·00	0·961	0·898
0·70	0·913	0·840
0·50	0·860	0·775
0·30	0·757	0·656

Table 1 presents a comparison of average ρ_{Mt}^2 with ρ_{MX}^2 for unstratified 18-item tests at various values of inter-item correlations. This table shows the magnitude of the difference. In general, as the item intercorrelation decreases, the magnitude of the difference increases. The same conclusion can also be made for tests constructed by stratification of item content, as can be seen from the data presented in Table 2. A comparison of ρ_{Mt}^2 and ρ_{MX}^2 for tests stratified only for content shows that test reliability for the unmatched item design is lower than the reliability of a comparable matched item design.

TABLE 2. COMPARISON OF MATCHED VERSUS UNMATCHED TEST EXPERIMENTS: TESTS STRATIFIED FOR ITEM CONTENT

(Data on ρ_{Mt}^2, α_s and α from Cronbach *et al.* (1965). γ_s and γ are averages based upon five samples of 200 cases each.)

r_w	r_b	ρ_{Mt}^2	α_s	α	ρ_{MX}^2	γ_s	γ
1·00	0	0·866	0·860	0·758	0·747	0·736	0·641
0·70	0	0·765	0·757	0·668	0·639	0·654	0·582
0·50	0	0·664	0·656	0·579	0·535	0·533	0·464
0·30	0	0·507	0·501	0·442	0·389	0·412	0·362

In addition, comparison of alpha and alpha-stratified coefficients for tests stratified for item content with the corresponding gamma and gamma-stratified coefficients in Table 2 shows that the discrepancy between gamma and gamma-stratified is of the same magnitude as the discrepancy between alpha and alpha-stratified. This, stratification for item content seems to have about the same effect for both the matched and unmatched item designs.

TABLE 3. COMPARISON OF MATCHED VERSUS UNMATCHED TEST EXPERIMENTS: SINGLE-FACTOR TESTS STRATIFIED FOR ITEM DIFFICULTY

(Data on ρ_{Mt}^2, α_s and α from Cronbach *et al.* (1965). γ_s and γ are averages based upon five samples of 200 cases each.)

r_w	ρ_{Mt}^2	α_s	α	ρ_{MX}^2	γ_s	γ
1·00	0·978	0·970	0·942	0·977	0·962	0·879
0·70	0·919	0·915	0·902	0·918	0·904	0·816
0·50	0·863	0·859	0·851	0·862	0·844	0·732
0·30	0·760	0·754	0·752	0·760	0·740	0·578

The results are quite different for tests stratified in terms of item difficulty as shown in Table 3. Comparison of ρ_{Mt}^2 with ρ_{MX}^2 shows that, for tests stratified for item difficulty, the test reliability using an unmatched item design is very close to the reliability of a comparable matched item design. Also, comparison of alpha and alpha-stratified with the corresponding gamma coefficients in Table 3 shows that gamma considerably underestimates ρ_{MX}^2 while alpha is a good estimate of ρ_{Mt}^2. The point to be made is that stratification for item difficulty seems to make little difference in the matched item design, but is all-important in an unmatched item experiment.

TABLE 4. COMPARISON OF MATCHED VERSUS UNMATCHED TEST EXPERIMENTS: TESTS STRATIFIED FOR BOTH ITEM DIFFICULTY AND CONTENT

(Data on ρ_{Mt}^2 and α_s from Cronbach *et al.* (1965). γ_s and γ are averages based upon five samples of 200 cases each.)

r_w	r_b	ρ_{Mt}^2	α_s	ρ_{MX}^2	γ_s	γ
1·00	0	0·926	0·912	0·935	0·924	0·712
0·70	0	0·787	0·780	0·789	0·794	0·527
0·50	0	0·675	0·669	0·676	0·695	0·402
0·30	0	0·512	0·508	0·513	0·510	0·126

Table 4 presents data for tests stratified for both content and difficulty. These results reinforce the conclusion that stratification for item difficulty is very important in an unmatched item experiment, and that under these conditions the reliability of unmatched item tests is quite comparable to the reliability of tests using the conventional matched item technique.

TABLE 5. ESTIMATED STANDARD ERROR OF γ AND γ_s AS A FUNCTION OF THE MODE OF STRATIFICATION, THE WITHIN-STRATUM INTERCORRELATION AND SAMPLE SIZE

Stratification	N	$r_w = 0{\cdot}30$		$r_w = 0{\cdot}50$		$r_w = 0{\cdot}70$		$r_w = 0{\cdot}99$	
		γ	γ_s	γ	γ_s	γ	γ_s	γ	γ_s
Difficulty	25	0·139	0·079	0·069	0·039	0·032	0·011	0·027	0·007
	50	0·101	0·050	0·027	0·014	0·032	0·016	0·028	0·006
	75	0·084	0·041	0·033	0·021	0·030	0·017	0·020	0·004
	100	0·044	0·022	0·026	0·015	0·013	0·009	0·019	0·004
	150	0·039	0·023	0·029	0·018	0·017	0·011	0·015	0·003
	200	0·030	0·020	0·025	0·015	0·011	0·008	0·009	0·002
Content ($r_b = 0{\cdot}0$)	25	0·153	0·138	0·179	0·186	0·049	0·046	0·049	0·039
	50	0·129	0·118	0·113	0·116	0·070	0·051	0·063	0·045
	75	0·119	0·106	0·079	0·078	0·063	0·042	0·054	0·036
	100	0·127	0·115	0·066	0·060	0·050	0·037	0·080	0·048
	150	0·102	0·094	0·035	0·031	0·049	0·041	0·070	0·044
	200	0·081	0·075	0·038	0·032	0·046	0·039	0·048	0·031
Content–Difficulty ($r_b = 0{\cdot}0$)	25	0·260	0·125	0·296	0·140	0·070	0·031	0·055	0·024
	50	0·152	0·069	0·150	0·070	0·094	0·037	0·035	0·010
	75	0·098	0·052	0·100	0·049	0·066	0·031	0·046	0·015
	100	0·059	0·036	0·093	0·049	0·064	0·032	0·044	0·014
	150	0·039	0·024	0·098	0·050	0·048	0·023	0·032	0·009
	200	0·007	0·009	0·071	0·033	0·041	0·019	0·018	0·007

Finally, turning to the data in Table 5 on the sampling variability of gamma and gamma-stratified, it is seen that the sampling variability of gamma-stratified tends to be smaller than the sampling variability of gamma for all stratification conditions—the discrepancy being most marked for small N and low item intercorrelations. However, it is clear that stratification by item difficulty produces the greatest effect on the sampling variability of gamma-stratified. These results reinforce the conclusion that stratification by item difficulty is extremely important when collecting unmatched data.

ACKNOWLEDGEMENT

Some of the data in Tables 1–4 are reproduced from the paper by Cronbach *et al.* (1965) with permission of *Educational and Psychological Measurement.*

REFERENCES

CRONBACH, L. J. (1963). Course improvement through evaluation. *Teacher's Coll. Rec.* **64**, 672–683.

CRONBACH, L. J., RAJARATNAM, N. & GLESER, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *Br. J. statist. Psychol.* **16**, 137–163.

CRONBACH, L. J., SCHÖNEMANN, P. H. & MCKIE, D. (1965). Alpha coefficients for stratified-parallel tests. *Educ. psychol. Measmt* **25,** 291–312.

LAWLEY, D. (1943). On problems connected with item selection and test construction. *Proc. R. Soc. Edinb. A,* **61,** 273–287.

LORD, F. M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika* **20,** 1–23.

LORD, F. M. (1962). Cutting scores and errors of measurement. *Psychometrika* **27,** 19–30.

OSBURN, H. G. (1967). A note on design of test experiments. *Educ. psychol. Measmt* **27,** 797–802.

RAJARATNAM, N., CRONBACH, L. J. & GLESER, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika* **30,** 39–56.

TUCKER, L. (1946). Maximum validity of a test with equivalent items. *Psychometrika* **11,** 1–13.

Vol. 21
Part 2
247–250

The British Journal of Mathematical and
Statistical Psychology

November
1968

A SCORING SYSTEM FOR MULTIPLE CHOICE TEST ITEMS

By J. O. Ramsay
McGill University

Weighted scores for the alternatives in multiple-choice items are proposed which maximize the separation in mean score between two criterion groups. These are the differences between the probabilities of choosing the alternatives for the two groups. Several advantages of this scoring system over the conventional ones are discussed and the method extended to the case of more than two criterion groups.

1. Introduction

The term ' multiple choice ' usually refers to test items in which there are more than two possible responses. For the purposes of this paper, the term will refer more broadly to any item with two or more possible responses, including items in which one response is correct and all other responses are treated collectively as the wrong one.

The almost universal method of scoring these items is the assigning of unity to the correct response and zero to all of the others. There are two much-discussed problems with this scheme. First, the mean score that can be attained by guessing is large, and the laws which guessing obeys are not generally known. The second and closely related disadvantage is that this scheme makes no provision for partial knowledge, for the fact that some alternatives are more wrong or right than others, or even that more than one answer may be correct. These difficulties have been discussed at some length by Chernoff (1962) and de Finetti (1965).

Although it is of obvious importance to arrive at a suitable *a priori* scoring scheme, much test development does or can take place on an *a posteriori* basis. That is, a pool of items may have been given to some body of examinees and it is desired to devise an efficient test on the basis of their responses. There is often some criterion available which can be used to divide the examinees into two groups. The test should then reproduce this division as well as possible so that the criterion need not be used on future groups.

2. Proposed Scoring Method

Suppose that the population of examinees can be separated into two groups (group 1 and group 2), that a total test score is indicated by T, and that this is a simple sum of the scores, S_i, obtained on n items. Suppose also that each of the k alternatives in item i has attached to it a weight, w_{ij}, and that the score on item i is simply the weight attached to the alternative which has been chosen.

If, in group r, the probability of choosing alternative j in item i is $P_i(j|r)$, then the mean item score is given by

$$E[S_i|r]=\sum_j w_{ij}P_i(j|r). \tag{1}$$

The fact that it is accepted practice in statistics to discriminate between two groups on the basis of their means suggests that the weights, w_{ij}, might be chosen so as to separate the means of the two groups by as much as possible. That is, they should maximize the following simple criterion:

$$Q=(E[T|1]-E[T|2])^2 \tag{2}$$

subject to the restriction:

$$\sum_i\sum_j w_{ij}{}^2=C. \tag{3}$$

If eqn. (2) is rewritten:

$$\begin{aligned} Q&=\left[\sum_i\sum_j w_{ij}P(j|1)-\sum_i\sum_j w_{ij}P_i(j|2)\right]^2\\ &=\left[\sum_i\sum_j w_{ij}\{P_i(j|1)-P_i(j|2)\}\right]^2; \end{aligned}$$

and it is noted that this really is the square of an inner product of two vectors with nk elements and therefore proportional to the square of the cosine of the angle between them, then it is apparent that Q will be maximized when

$$w_{ij}=\alpha[P_i(j|1)-P_i(j|2)] \quad (\alpha\neq 0). \tag{4}$$

Two useful consequences of choosing the w_{ij} in this way are that the weights for a particular item sum to zero and that the sum of the squares of the weights for a particular item have a lower limit of zero and an upper limit of $2\alpha^2$. This can be seen from the following argument:

$$\begin{aligned} \frac{1}{\alpha^2}\sum_j w_{ij}{}^2&=\sum_j[P_i(j|1)-P_i(j|2)]^2\\ &=\sum_j P_i(j|1)^2+\sum_j P_i(j|2)^2-2\sum_j P_i(j|1)P_i(j|2)\\ &\leqslant\sum_j P_i(j|1)^2+\sum_j P_i(j|2)^2\\ &\leqslant\left[\sum_j P_i(j|1)\right]^2+\left[\sum_j P_i(j|2)\right]^2.\\ &\leqslant 2. \end{aligned} \tag{5}$$

This latter result is of importance since $\sum_j w_{ij}$ is a measure of how well item i separates the means of the two groups and hence valuable for item analysis purposes.

Argument (5) also implies that the maximum separation of the two groups by an item occurs only when all the members of one group choose one alternative and all the members of the other group choose one other alternative. That is, the best multiple choice item in this sense is really a two-choice item.

Criterion (2) is a special case of a more general criterion applicable to g groups with criterion scores X_r $(r=1, \ldots, g)$. If the X_r's are scaled so that $\sum_r X_r = 0$, then maximizing the following criterion is equivalent to maximizing the covariance between the criterion scores and the expected test scores:

$$Q = \left\{ \sum_r^g X_r E[T|r] \right\}^2. \tag{6}$$

If this criterion is rewritten:

$$\begin{aligned} Q &= \left\{ \sum_r^g X_r \sum_i^n \sum_j^k w_{ij} P_i(j|r) \right\}^2 \\ &= \left\{ \sum_i^n \sum_j^k w_{ij} \left[\sum_r^g X_r P_i(j|r) \right] \right\}^2, \end{aligned} \tag{7}$$

then it is again apparent that the optimal weights are given by:

$$w_{ij} = \alpha \sum_r^g X_r P_i(j|r). \tag{8}$$

These weights are identical to those given by eqn. (4) when $g=2$, except possibly for a scale factor.

3. DISCUSSION

There are several advantages to these scoring schemes aside from those mentioned above. The substitution of sample estimates of the probabilities in eqns. (4) or (8) provides minimum variance unbiased estimators for the weights. These weights are especially useful for items of personality inventories used for diagnostic purposes in which the alternatives have no clear *a priori* weightings. For the more common tests in which one answer is correct, these weights give an advantage in mean score to those who have partial knowledge and give to those who guess randomly amongst all the alternatives a mean score of zero. Faking on a test of this kind could give a significantly large negative total score. Finally, because partial knowledge is rewarded, the reliability and validity of a test scored in this manner should be greater than for a test scored by the 0–1 scheme.

These weights do have two minor problems. The first concerns criteria (2) and (6). Neither of these implies the minimization of misclassifications; and if a situation arises in which the increase in variance of the scores in the two groups has an effect which exceeds the effect of separating the means, then little may be gained by this criterion. However, it seems unlikely that this will often arise in practice.

The second problem is the danger of perpetuating bad or out-of-date criteria for dividing the population. To some extent this difficulty is common to all item analysis procedures, and whether or not it would be more acute in this situation is not clear. This underlines the necessity of continually updating test criteria.

In summary, if the weighting of an alternative is made proportional to the difference between the probabilities of choosing this alternative for two criterion groups, then this is a scoring system which not only involves a bare minimum of computation and lends itself easily to item selection, but it will also generally distinguish between the two groups efficiently and improve the reliability of the test score.

REFERENCES

CHERNOFF, H. (1962). The scoring of multiple choice questionnaires. *Ann. math. Statist.* **33**, 375–393.

DE FINETTI, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *Br. J. math. statist. Psychol.* **18**, 87–123.

Vol. 21
Part 2
251–263

The British Journal of Mathematical and
Statistical Psychology

November
1968

ATTRIBUTE AND RELIABILITY

By John Ross and James Lumsden
The University of Western Australia

An analysis of classical test theory and of more recent developments shows that the concepts of true score, parallel test and reliability are poorly defined. Where reliability is defined as generalizability it is not so much poorly defined as detached from the essential interior properties of the test. It is proposed that tests as such be restrictively defined as devices which map an attribute in a single dimension. A logistic model which fits this conception is examined and it is shown that true score, parallel test and reliability can be defined without ambiguity, inconsistency or circularity. Deeper investigation reveals that the true score concept is not only unnecessary as a basis of reliability but also a distraction. Reliability is formulated as a characterization of the test's mapping. It is shown that the mapping can be characterized for the model and it is also shown that there is a distinction between information loss due to regression and error which probably lies behind the poorly conceived notion of true score.

1. Contemporary Views of Reliability

The reliability problem is to determine the extent to which the estimates given by a test might be in error. There have been numerous attempts at solution, tracing back to the work of Spearman (1910). It is shown in this paper that attempts so far have been unsuccessful, although they have produced solutions to other related problems. A solution to the problem is then presented by restricting the problem to tests which can be considered to be devices that measure an attribute. The solution is illustrated by a method of describing reliability fully for itemized tests which meet the requirements of a logistic item characteristic curve model.

First, an examination is made of three relatively recent positions which, between them, span the range of modern thinking on the subject.

Classical test theory

Classical test theory, by which we roughly mean test theory as codified by Gulliksen (1950), bases its approach to reliability on a theoretical analysis of tests. It constructs a simple model using the theoretical concepts of true score (T), parallel forms (PF) and error (E), in terms of which observed scores and the variation in them can be explained. Observed score variance is said to be the sum of true score variance and error variance. Reliability is defined in terms of true score variance and error variance, and is estimated from the correlation between parallel forms. In terms of the model it can be shown that the reliability coefficient r_{xx} is given by

$$r_{xx} = \sigma_t^2/(\sigma_t^2 + \sigma_e^2),$$

where σ_t^2 is true-score variance and σ_e^2 is error variance.

Two of the three theoretical terms, T and E, are tied directly to observed score X, by the formula $X=T+E$. The third theoretical term, PF, is defined in terms of T and E by the following conditions:

(i) $T_{ig}=T_{ih}$ ('The true score of any person on one test must equal the true score of that person on the other parallel test'—Gulliksen, 1950, p. 11).

(ii) $S_{eg}=S_{eh}$ ('For two parallel tests, the errors of measurement are equal'—Gulliksen, 1950, p. 12).

It is impossible to apply the classical model as it stands, since parallel forms are defined in terms of T and E, to which there is no direct access, with the result that observed variance cannot be decomposed into the required theoretical variances of σ_t^2 and σ_e^2.

An application of sorts is achieved by buttressing the model with a set of criteria for parallel forms. A set of tests is said to be parallel (Gulliksen, 1950, ch. 3) if

$$\text{(i)}\quad \bar{X}_1=\bar{X}_2=\ldots=\bar{X}_k;$$
$$\text{(ii)}\quad S_1=S_2=\ldots=S_k;$$
$$\text{(iii)}\quad r_{12}=r_{13}=\ldots=r_{jk}.$$

Gulliksen later adds that the validities should be equal for any criterion. If the requirements are met, it can be shown that r_{gh} (the correlation between any pair of tests within the set) is equal to r_{xx} or $\sigma_t^2/(\sigma_t^2+\sigma_e^2)$. Furthermore, the true score of T_i of any individual T_i can be expressed as a limit

$$T_i=\lim_{K\to\infty}\sum_{g=1}^{K}(X_{ig})/K.$$

However, all obstacles to the application of the model are not removed by the criteria. Given sufficient tests, we can always find a small subset, say two or three in number, with the same means and variances which meet the correlational criteria. But without criteria as to what constitutes a sufficient or an appropriate set we could never be sure if the tests in the subset were really parallel or whether we had so restricted the subset that non-parallel tests appeared parallel. There needs to be some rule of test selection, independent of the statistical criteria, to guard against capitalizing on sampling fluctuations of test statistics. It has been suggested that expert judgement on content might define the set, but this leaves Guttman's (1953) objection that more than one equally acceptable basis may be found for forming a set, which would mean that a test would be parallel in different sets, and hence have different reliabilities. If so, it would also have a variable theoretical constitution, in terms of T and E, depending upon shifts in expert opinion.

The fundamental deficiency in the classical analysis is not simply the lack of any basis within the definition of parallel test to select the sample of tests to which the statistical criteria are applied, but the lack of any basis in the theory for the formation of the set. It is necessary to have such a basis, and to have it in the theory if the theory is to provide a genuine foundation for the analysis of reliability.

Some theorists have been drawn to the conclusion that the reliability question must be re-phrased as a question about generalizability. We will consider this view in the next section, and show that it amounts to abandoning the problem of reliability in its original form.

Reliability as generalizability

Cronbach *et al.* (1963, p. 144) have redefined the problem of reliability: 'An investigator asks about the precision or reliability of a measure because he wishes to generalize from the observation in hand to some class of observations to which it belongs.' They pin reliability not to a test, but to the universe of conditions over which an investigator wishes to generalize. There are thus as many reliabilities for a test as universes that investigators are able to specify.

The position reached by Cronbach *et al.* is an important generalization and a natural extension from previous theoretical studies. The authors say their rationale is one which combines numerous lines of thought about reliability and allows reinterpretation of them.

The 'liberalization' of reliability theory that Cronbach *et al.* offer is valuable for a variety of practical purposes. In a given administrative context, there might be a well-defined universe of conditions (such as the raters available, the alternative test forms available, or the testing locations available) and the investigators may well wish to know about variance in estimates with variation in any or all of such 'facets'. But the investigator who requires a theoretical analysis of his test, one such as classical theory attempted to provide, will not find it.

Cronbach *et al.* present a model which might be taken to be the basis for a theoretical analysis. The model accounts for scores obtained by an individual in any condition of the universe in terms of a universe score and several subsidiary scores. The universe score is defined as the individual's score on the first centroid factor of the covariances between the conditions in the universe. The subsidiary scores are scores on other centroid factors required to account for the residual covariances, and an error score.

It is shown that, from the model, it is possible to estimate the correlation between universe score and obtained score provided three conditions are met: '(1) The universe is described unambiguously, so that it is clear what conditions fall within the universe. (2) Conditions are experimentally independent... (3) Scores X_{pi} are numbers on an interval scale.'

The decomposition of obtained score variance into universe score variance and other variance might be taken as a sufficient theoretical analysis, substituting adequately for the classical target of a true variance, error variance decomposition. The difficulties with such an interpretation are as follows. Defining the universe score as a factor score on the centroid factor from a covariance matrix makes universe score change with changes in the scaling unit of any test in the universe (see Ross, 1963). Secondly, if the centroid is to be interpretable from the definition of the universe, the universe must be defined in theoretical terms, and the structure of the test population must be such that the first centroid factor corresponds to the factor common to all the tests, and is not contaminated by secondary factors. It is difficult to see that such conditions could be met without a very precise and complete universe definition, and without a test construction theory sufficient to ensure that the centroid corresponded with the definition. Universe score variance will not, except in most exceptional circumstances, correspond to any theoretical score, or at least to one which the investigator can characterize before he selects the sample of tests for factor analysis.

As we posed the reliability problem, it was to determine the extent to which the estimates given by a test might be in error. The approach through generalizability leads to a multiplicity of answers, or, it might be better to say, to a different question, not now about a test, but about the set of universes to which it may be regarded as belonging. The original question remains unanswered, and, as we have seen, there is no easy answer to it by appealing to the concept of universe score. An attempt will be made to show that the original question is answerable by restricting it to tests which can be considered to be measures of an attribute.

Modern test theory

Modern test theory is used here to designate test theory as presented by Lord (1952) and as developed by Lord himself and others. Modern test theory differs from classical theory more than is commonly recognized; these differences have been minimized in several presentations and applications. First, it works from an attribute or a universe of content. Secondly, it deals usually only with tests of one attribute or one universe of content. Thirdly, it need not define the concept of error directly. Fourthly, it does not speak of true score in the same sense as classical theory. Fifthly, it deals explicitly with itemized tests, which classical theory does not. Sixthly, in virtue of this, it does not treat raw test score as *necessarily* the fundamental datum of a test.

Although the fundamental monograph is Lord's *A Theory of Test Scores* (1952), a more influential statement of the modern test theory position on true score is that given some years later (Lord, 1959). It might be noted that the 1959 position is nearer the classical than is the 1952. Lord gives three approaches to the problem of making inferences about true score. All assume that true score differs from observed score only by error of measurement and that the expected value of error is zero. The first is to assume that it is possible to obtain two tests on which scores differ only because of error of measurement. The second is to define true score as hypothetical score on a test made up of a large pool of items which have properties the same as those which make up the test at hand. A variant of this approach is to assume that the test is a random sample of items from such a pool. The third approach makes assumptions about the distribution of the error of measurement components of observed score around their expected value of zero. Lord goes on to make distinctions (lacking in classical theory) between the true score of a test and the ' ability underlying the test ' or latent continuum. True score normally has a limited range, whereas the latent continuum has not. More importantly, ' non-parallel tests of the same psychological dimension ' will each necessarily have a different true score metric but the same latent continuum (or underlying ability metric).

Any test thus has three ' metrics ', the observed score metric, the true score metric, and the latent continuum metric. Observed score is related to true score by error, and true score is a monotonic, but non-linear transformation of the latent continuum.

The point made about the true score models is that they make weak assumptions, in contrast to latent continuum models which make strong assumptions about item trace lines (cf. Lord, 1952). Some information is lost by settling for a weak model, since one works from true score which has a perfect but *curvilinear* correlation with the latent continuum. Lord's argument appears to be that the price of information loss is worth paying for the gain in applicability of the true score models.

The conceptual difficulties of the classical approach are not removed in the various true score models. The matched-forms model encounters Guttman's objections about the multiple bases for the formation of parallel test sets. The item-sampling model simply removes the same problem to the item level, and raises the further problem of attaching true score or reliability to any *particular* test rather than to some imperfectly defined set of tests of which the particular test of interest is a member. The model or models which make assumptions about the distribution of errors are conceptually identical to the classical model and meet the same objections.

In Lord's (1952) monograph, true score held a less central position. Each item in a test was tied to a common attribute by an item characteristic curve. A true score could be and was defined in terms of the model, not as a primitive term, and error similarly could be defined in terms of the model. The model was testable in terms of predictions it made from item parameters about the shape of obtained score distributions, and it provided, in principle, a basis for inference from the vector of item responses to location of an individual on the attribute. The difficulty with Lord's model was the intractability of its mathematics, because of the assumption of the normal ogive as the form of the item characteristic curve. The model itself, however, is theoretically sound, and allows a coherent treatment of the reliability question in its original form. This treatment does not involve a consideration of true score.

It will be shown that a slight modification of Lord's model, made by Birnbaum (1957) which removes the mathematical difficulties, provides one class of test for which it is possible to give a coherent treatment of reliability.

2. AN ATTRIBUTE-BASED ITEM MODEL

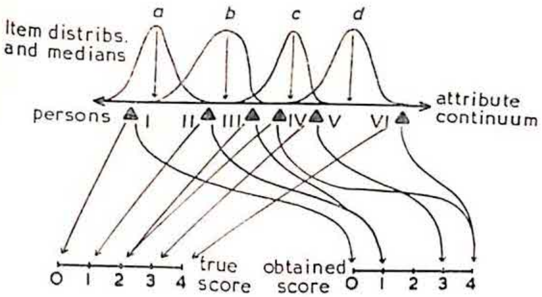
To accept an attribute model like Lord's, or the Birnbaum revision which is to be presented, is to admit that there is a class of tests and a class of non-tests; that there may be procedures which look like tests, but which are not. An attribute-based test can be depicted theoretically as follows. There is an attribute space, a metric space which supports variation in the attribute. The test itself *maps* the attribute space into a test space, which is usually the real numbers or the integers, but which may be something else like a category system or a multidimensional metric system. A test may be developed in the belief that there is a well-defined attribute which it is mapping, but the belief may be false. If so, the test is not a test at all in our restricted sense. Most commonly, itemized tests comprise items which are not homogeneous, which is to say that the mapping they produce is governed by several attributes at once. While it can never be finally proved that a test is a test in our restricted sense, it is possible to prove that a test is not a test of any attribute, since a mininum requirement is unidimensionality, which can be checked in several ways (Lumsden, 1957). Birnbaum's model provides a second check, since the assumption that the attribute governs the response to each item allows a precise prediction about the distribution of test scores. If the obtained distribution is not as predicted, we must reject the test if we accept the model.

The attribute model for itemized tests (called model A, to distinguish from the true score–error score model, or model T) may be sketched as follows.

The basic model first supposes that there is an attribute which is to be measured, and which may be represented as a directed continuum. Second, it supposes that persons each have a point location on the continuum. Thir it supposes that items each have a distribution of point locations on th continuum. A person *truly* passes an item if his location is above that of th median for the item, and truly fails if it is not. Since continuity is assumed the possibility of a person having the same location as an item may be ignored.

If true score is needed, it may be defined for a given sample of items (not a universe of items), as the number of items which a person truly passes. The basic model is represented in Fig. 1, which, shows the attribute continuum, the location of a set of persons and items, and separately, the true score of the persons, on a separate continuum. The figure also shows obtained score as independently derived from relationships on the attribute continuum and not from true score. It is to be noted that some information is lost in the passage from location on the theoretical attribute continuum to true score, and that the true score–theoretical attribute relation is not linear. It departs from linearity by virtue of the upper and lower bounds upon true score, in virtue of the discontinuities in true score, and in virtue of the fact that true score depends upon item locations as well as person locations on the fundamental continuum.

The point location of items and persons is to be interpreted as a median of a distribution of locations. Both items and people 'wobble', and the frequency with which they appear in locations along the continuum can be described by a probability density function. At this stage no particular function is specified, although the normal density function and the logistic density functions are commonly postulated. It is further assumed that people shift location only slowly over time, but that items shift location instantaneously. This is tanta-



Individual	*True score*	*Obtained score*
I	0	0
II	1	1
III	2	1
IV	2	4
V	3	3
VI	4	4

FIGURE 1. The relation between the attribute continuum, true score and obtained score. Persons are shown as having point locations (triangles below the top line) and items as probabilistic locations (probability density distributions above the top line) along a directed attribute continuum. True score and obtained score are shown as independently derived from the attribute locations of persons and items. True scores are determined from medians of item locations, a person passing an item if his location is to the right of its median location. Obtained scores are determined from the probability density distribution. The probability of a person's passing an item is the probability under the density distribution for the item to the left of the location of the person. Some possible discrepancies between true score and error score are shown in the table inset.

mount to assuming that there is an indeterminacy in the location of items which s not the case for people. Since it is assumed that a person passes an item at any esting if his location is above the location drawn from the item distribution, btained score is thus undefined except in probabilistic terms. True score is defined at any given instant, but may vary over time. The question of attribute stability is thus made quite distinct from that of attribute measurement. It might be noted that it makes no essential difference to locate the momentary indeterminacy in the persons, or to spread it over both persons and items. It has been chosen to locate the indeterminacy in the item in order to show that it is possible to construct a coherent theory on the assumption that persons have a fixed location, which is equivalent to assuming that they have a definite amount of the attribute.

Where true score is defined in terms of the momentarily fixed person location and the *median* of the momentary distribution of item locations, obtained item score is defined in terms of the fixed person location and a random location from the item distribution. True score and obtained item score are defined independently in terms of the attribute and the characteristics of items and persons. Obtained score is not defined directly as true score plus error. This is a fact which *follows from* the definition of true score and obtained score. There are consequences of considerable importance when we come to consider variation in error over the true score range.

True score is defined in terms of a fixed set of items, but a person's location on the attribute is independent of any set of items which might be devised. Thus as item sets vary, true scores will vary, and one can ask about any set of items, how well their true score represents the attribute continuum. Pairs of item sets may consist of items which are matched one-for-one in median location, but the distributions of the items may show greater dispersion in one set than in the other. Such pairs of sets would yield the same true score, but the obtained score would vary about the true score more in the one case than the other. If we wish to define parallel forms of a test we may do so by defining them as tests which yield the same true score for all individuals. One could also define *matched* forms of a test, which are both parallel, and have items matched for dispersion as well as median location. Once these distinctions are made, it is clear that there are three questions in the area usually called reliability: (1) How well does the true score represent the attribute? (2) How well does the obtained score represent the true score? (3) How well does the obtained score represent the attribute? The last is the central question.

Other types of test

So far, only itemized tests have been considered. The axiomatic analysis of measurement in psychology, notably by Suppes & Zinnes (1963), and others since, has made it clear that some empirical or data structures permit the inference that properties of objects can adequately be represented by numbers, and that others do not. That the inference can be drawn can be shown by the proof of a representation theorem, which establishes that there is a functional mapping from the empirical structure to the real numbers. Any procedure for which the existence of a mapping function could be established would be a test in our sense.

Most cases studied by axiomatic means have employed the direct comparison of objects one with another, as in the balance test for weight, or as in choice tests for preference. It seems to have become uncritically accepted that only such direct comparison provides a proper basis for measurement. The attribute model for items which we propose provides *indirect* comparison.

While no axiomatic treatment of the model proposed as an example of a proper test has been undertaken, it seems certain that a satisfactory treatment could be developed, assuming the existence of a sufficient supply of items for any given attribute. A transitivity axiom for ordering could be developed from the requirement that if A is scaled higher than B and B higher than C on one sample of the items, then A will be scaled higher than C on other samples of sufficient size. Different samples would provide indirect but unprejudiced comparisons of subjects. Similarly, transitivity of scaled score differences could be established from the empirical differences in scaled scores in different item samples.

3. The Logistic Version of an Attribute-Based Item Model

In Birnbaum's (1957) version the item characteristic curves are logistic in form. The logistic function, like the normal, is completely specified by two parameters: one specifying location, the other dispersion. Birnbaum shows that both location *on the attribute* and dispersion may be estimated if both the correlation of the item with the attribute and the difficulty of the item are known, assuming one knows the distribution of individuals on the attribute. A test which meets the Birnbaum model formally satisfies the requirements of a testing device as we have developed it, and it is possible to estimate the parameters needed to apply the model. A strong test of the model is possible for any given test. It is possible to derive a predicted distribution of a weighted test score from the item parameters alone. Ross (1966) has recently shown that the model fits some six tests varying in content, item format, item difficulty and average inter-item correlation.

From the same item parameters which yield the predicted score distribution one can also compute a regression of test score on the attribute and the variance of test score about the regression as a function of the attribute.

An example of the application

Birnbaum's model assumes that there is a single ability involved, Y, and that the probability, p_g, of passing item g is given by the logistic function. The basic equation in the model is thus

$$p_g = 1[1 + \exp(b_g - a_g y)],$$

where a_g is a dispersion parameter depending upon the correlation between item g and the attribute, and b_g is a location parameter. The definitions of a_g and b_g are

$$a_g = C r_g \surd(1 - r_g^2), \qquad b_g = N^{-1}(p_g) \surd(C^2 + a_g^2),$$

where C is a constant (approx. 1·7), r_g is the item–attribute correlation, and $N^{-1}(p_g)$ is the normal deviate corresponding to the proportion of the population passing the item.

With these definitions, and by methods not described here (see Birnbaum, 1957; Ross, 1966), Birnbaum develops a weighted test score, t, which is defined as the sum of the items passed, each weighted by a_g, the item parameter defined above. Birnbaum then develops three propositions which are central to our development: (1) that the statistic t is a sufficient statistic, i.e. it preserves *all* the information in the test protocol relevant to the estimation of the value of Y; (2) that there is a simple equation, using a_g and b_g, which gives the expected value of t for any value y of Y; (3) that there is another simple equation, again using a_g and b_g, which gives the asymptotic variance of t for any value y of Y.

The direct implication is that if the model fits, a maximally efficient procedure can be set up for estimating the attribute, not from observed score but from the weighted score t, which is the sufficient statistic. Furthermore, the model is self-protective in that if a distribution of ability is assumed, and if a_g and b_g are estimated for each item, it can then be estimated what the distribution of t scores should be. If they are as predicted, there is solid supporting evidence that the model fits and that the test is a test in the general sense of a device which maps one attribute into a test score. One estimation procedure is to use first factor loadings as estimates of r_g and proportion of a simple passing as estimates of the population proportion.

The two parameters of the logistic model were estimated for two multiple-choice verbal tests and four open-ended number tests by Ross (1966). A normal distribution on the attribute was assumed. The expected and obtained distributions for the two of the six tests studied are shown in Fig. 2. The distributions for the other four tests together with a full description of the procedure are given by Ross (1966). The model was also used to obtain the regression of the sufficient statistic on ability and the variance of the sufficient statistic as a function of ability. Both functions for each test are shown in Fig. 3. The point which is at

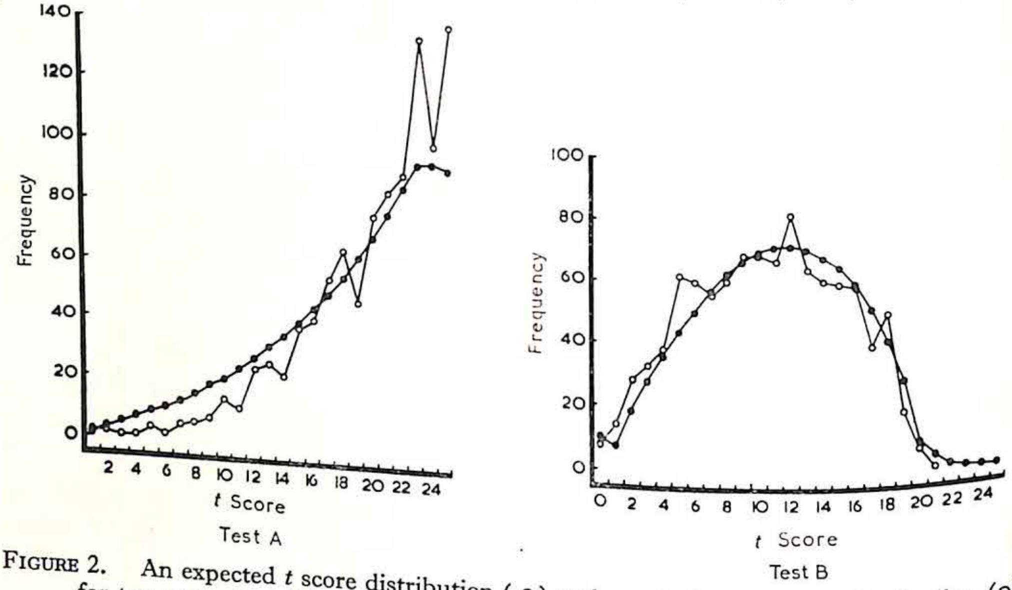


FIGURE 2. An expected t score distribution (●) and an obtained t score distribution (o) for two tests, A and B, each somewhat unusual ($n=1{,}000$). For details, see Ross (1966).

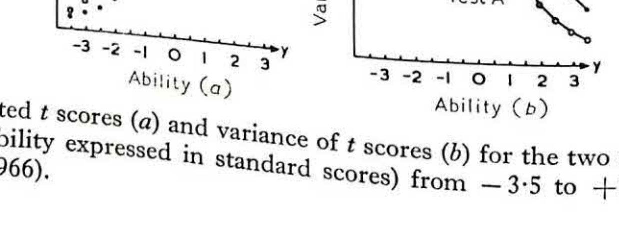


FIGURE 3. Expected t scores (a) and variance of t scores (b) for the two tests A and B, for y values (ability expressed in standard scores) from $-3{\cdot}5$ to $+3{\cdot}5$. For details, see Ross (1966).

once obvious is that both the regression function *and* the variance function are non-linear, which makes it impossible to reduce the relation between the statistic and the attribute to any single index. The average variance could be computed, or an average weighted variance according to the attribute distribution, but this would not be helpful. The fit between the expected and obtained distributions encourages the belief that the model is close to being realized in the two tests studied. Even better fits were obtained with four other tests which gave extreme U-shaped distributions.

Implications for reliability

The reliability problem, as stated, is to determine the extent to which estimates given by a test might be in error. For attribute-based itemized tests, which map from attribute to test space, the problem is to describe precisely the error or variability in the mapping. Birnbaum's model first describes the mapping by way of a regression of test score on the attribute (see Fig. 3), and secondly describes the error in the mapping by giving the variance of test score about the regression as a function of the attribute. This is a complete answer to the reliability problem as it was posed. The answer is not in the form of an index, but of two functions which together provide the required description of the test mapping. Indices comparable to reliability coefficients might be derived, but they would be incomplete descriptions in all but very special cases. An answer in similar form is possible in principle for tests of all types, not only for itemized tests.

4. THE FALSE TRAIL OF TRUE SCORE

In his 1952 monograph Lord scarcely mentioned true score, nor needed to for his development of test theory. In 1959, however, he called true score the fundamental concept, and treated it as a genotype or ideal, an obtained score stripped of error. The shift in Lord's position has been accompanied by a concern to find models (like the item-sampling models) which will fit testing situations which do not satisfy the strict demands of the 1952 model. In cases where it makes no good sense to speak of an attribute, it is tempting to set up true score as an alternative ideal. But, as argued above, true score needs to be defined from an attribute and will not stand, by itself, as an *interior* property of the test. The outcome of the attempt to set up true score as an ideal is, in point of fact, the position reached by Cronbach *et al.*, which attributes to the test as many true scores as universes which may be found for it. In brief then, while true score may suffice as a conceptual tool with which to examine the exterior properties of a test (its relations to other tests and criteria) it will not suffice for examining the intrinsic and singular interior properties of the test. The hard truth, as insisted earlier in this paper, is that there do exist tests in the usual sense which cannot be handled as tests in the psychometric sense. Some tests may be separable into their unitary components, as a mixture is separable into components, but only thus can they be dealt with as psychological measuring devices.

The attribute is needed to define true score. The point which next emerges is that if one has items related to an attribute as described in the Birnbaum model, the concept of true score is not needed. The model defines a sufficient statistic based on item weights which in turn depend (inversely) on the item dispersion. The model further defines regression of the sufficient statistic on the attribute—the expected value of the statistic for any attribute value—and a variance value for the statistic at any point on the attribute. Thus it is possible to set up the bivariate surface and the other regression of attribute on the statistic, and so to set up a procedure which directly estimates the attribute from the test protocol via a single value which carries all the relevant information the test protocol has to give. In such a system, what might true score be? It can hardly be the attribute since the two are not the same. On the other hand, it cannot be the sufficient statistic stripped of error. Error, or the fact that items have a probability slope on the attribute, is what makes it possible to make estimates of location along the attribute continuum. In the errorless case—the case of perfectly reliable items, which is the other possible interpretation of the true score—the sufficient statistic becomes undefined since the item weights tend to infinity. It is tempting to imagine that the true value of the statistic is its expectation over an infinite set of matched tests. If such a ' true score ' were obtained, it would not lead to the same estimate of attribute value as the expectation of the attribute values estimated separately from each test, unless the regression of ability on the statistic were linear, which can never be the case.

In the Birnbaum model, to seek true score is to seek a will o' the wisp, and it is suggested that psychometric theory followed a false trail when it turned away from the problem, tackled by Lord in 1952, of estimating the value on the attribute from the test protocol. Whenever a procedure is set up which does estimate the attribute value, it will be unnecessary and misleading to think of an ideal value of the estimator.

5. Test and Attribute: Information and Error

The position may be reviewed as follows. To make sense of the reliability problem, the test was considered as a device which maps from the attribute continuum into the test space, and, on this basis, it was discovered that the concepts of true scores or parallel tests are not needed in order to characterize the mapping. It is maintained that the reliability problem is to characterize the mapping, and that although it is possible to make sense of the classical concepts of true score, parallel test and reliability for item counts it is a distraction to bother to do so.

There is a distinction, which may be made and which may be conceptually valuable, between information loss and error in relation to the mapping process. The distinction is based on the fact that it is possible in mapping to lose information without being involved in error, by virtue of a many to one reduction

Error, considered as noise, constitutes a further source of information loss, but of a different kind. One may speak of two tests with a similar regression of attribute on statistic as informationally similar, in the sense that both are bound by the same informational limit, and regard the test with the greater conditional variance about the regression as differing in the second source of information loss, error. It is probably this distinction which lies at the heart of the true score concept which leads to the so-called attenuation paradox. The distinction seems to serve no practical purpose in the case of strict measurement, i.e. location along an attribute. It is, however, conceivable that one may wish to map into other sorts of spaces (which is not strictly measurement) and that it will be helpful to remember that the test space and the attribute space must correspond in certain ways, as well as that information loss and error must be low if the test is to give the mapping required.

REFERENCES

BIRNBAUM, A. (1957). On the estimation of mental ability. *Probability and Statistics in Item Analysis and Classification Problems.* Technical Report no. 15. School of Aviation Medicine, USAF.

CRONBACH, L. J., RAJARATNAM, N. & GLESER, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *Br. J. statist. Psychol.* **16,** 137–163.

GULLIKSEN, H. (1950). *Theory of Mental Tests.* New York: Wiley.

GUTTMAN, L. (1953). A special review of Harold Gulliksen: ' Theory of Mental Tests '. *Psychometrika* **18,** 123–130.

LORD, F. M. (1952). A theory of test scores. *Psychometric Monogr.* no. 7. Chicago: University of Chicago Press.

LORD, F. M. (1959). An approach to mental test theory. *Psychometrika* **24,** 283–302.

LUMSDEN, J. (1957). A factorial approach to unidimensionality. *Aust. J. Psychol.* **9,** 105–111.

LUMSDEN, J. & ROSS, J. (1963). Validity as equivalence. (Unpublished manuscript.)

ROSS, J. (1963). The relation between test and person factors. *Psychol. Rev.* **70,** 432–443.

ROSS, J. (1966). An empirical study of a logistic mental test model. *Psychometrika* **31,** 325–340.

SPEARMAN, C. (1910). Coefficient of correlation calculated from faulty data. *Br. J. Psychol.* **3,** 271–295.

SUPPES, P. & ZINNES, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology,* vol. 1. New York: Wiley.

THORNDIKE, R. L. (1951). In E. Lindquist (ed.), *Educational Measurement.* Menasha: Banta.

NOTES AND CORRESPONDENCE

A NOTE ON AN EXCEPTION TO THE RULE OF DECREASING EIGENVALUES IN SUCCESSIVE FACTOR ITERATIVE METHODS

By H. F. Priest
University of Canterbury, Christchurch, New Zealand

In basic structure successive factor methods it is assumed that eigenvalues will be found in descending order of magnitude. However, this is conditional not only on a sufficient number of iterations being carried out, but also on there being no vector in the matrix which is orthogonal or nearly orthogonal to the trial vector. An actual case is quoted.

As larger computers are brought into use, successive factor iterative methods may become obsolete. However, as of now there are plenty of smaller machines in use and, at least for didactic purposes, the method may be with us for a long time yet.

In basic structure successive factor methods, it is customary to set up a first trial vector which has all values at unity. It is then assumed that successive eigenvectors will have eigenvalues in decreasing order of magnitude. The ‘ root one ’ criterion for stopping the extraction of factors—widely used despite objections (Horn, 1966)—depends on the truth of this assumption. So, of course, does any other such criterion. We are warned that this assumption is conditional, e.g. ‘ This is true in the exact mathematical sense only if the iterations for each vector have been carried far enough ’ (Horst, 1965, p. 158). The assumption, however, carries with it another condition. The other condition is that there is no vector in the matrix which is orthogonal or very nearly orthogonal to the trial vector. When this second condition is violated the eigenvalues will not necessarily be produced in descending order of magnitude.

A particular matrix of correlations (which arose from a matrix of real data collected in a clinical setting) is as in Table 1.

Table 1. Correlations

1							
2	58						
3	−11	−25					
4	76	85	−23				
5	75	66	−27	85			
6	67	88	−17	65	61		
7	87	87	−22	96	80	76	
8	−11	−25	−11	−23	−27	−17	−22

The curious relation of variables 3 and 8 is the point of interest here.

Horst's program 7.4 (1965, p. 608) ‘ Solution Without Residuals ’ is used to demonstrate the point. If the above matrix is factored with this program (making only those amendments necessary for it to run on an IBM 1620) it produces two vectors with eigenvalues 4·98 and 0·820. Having found an eigenvalue less than one, it stops. If the program is then amended to extract further vectors, it produces eigenvalues 4·98, 0·820, 1·111, 0·570, 0·299, etc. *in that order*. If the program is again amended, so that it sets up a first trial vector $A(I, L) = I$ (effectively 1, 2, 3, 4, etc.), then it produces the same vectors and eigenvalues, but this time in descending order of magnitude of eigenvalues.

To show that number of iterations is not relevant here, the matrix in Table 2 was produced using Horst's program 7.4 (*op. cit.*), with a variable stop amendment, the unit trial vector, a comparison value (P) of 0·00000001, allowing up to 40 iterations and eight decimal places.

TABLE 2. EIGENVECTORS, EIGENVALUES AND ITERATIONS

	1	2	3	4
	0·845	0·282	0·000	−0·292
	0·901	−0·045	0·000	0·386
	−0·275	0·595	−0·745	0·021
	0·942	0·010	0·000	−0·144
	0·873	−0·081	0·000	−0·323
	0·836	0·119	0·000	0·454
	0·975	0·093	0·000	−0·052
	−0·275	0·595	0·745	0·021
Iterations	6	14	26	16
Latent roots	4·983	0·820	1·111	0·570

As can be seen, vector 3 is perpendicular to the trial vector of ones.

The matrix of correlations is admittedly unusual, but it does arise from real data. It may be argued that it should not have been factored, but that is not really the point at issue. The point is that the 'wrong' order of eigenvalues can occur in a real situation.

REFERENCES

HORN, J. L. (1966). A note on the use of random variables in factor analysis. *Br. J. math. statist. Psychol.* **19**, 127–129.

HORST, P. (1965). *Factor Analysis of Data Matrices.* New York: Holt, Rinehart & Winston.

BOOK REVIEW

Timing in Temporal Tracking. By J. A. MICHON. Soesterberg: Institute for Perception, RVO-TNO, 1967. Pp. 127.

In this monograph Michon discusses a series of experiments and develops a model accounting for the accuracy with which subjects can tap either at a given rate or in synchrony with a train of clicks. As such it would appear to be only relevant to psychologists whose research interests already include this narrow area. This would, however, be a very superficial assessment, for the monograph deserves a much wider readership.

Michon sees his experimental paradigm as a microcosm in which mechanisms important in a wide range of behaviours can be more easily studied in detail; in particular those mediating the exact timing of actions, and which are now considered essential in both psycholinguistics and motor skills. The relation between this experimental microcosm and more complex behaviours is clearly discussed, a welcome change from the common strategy of some psychologists, of apparently taking for granted the value of studying a certain microcosm, simply because other psychologists have studied it.

One might expect the author, in traditional fashion, simply to state his model of how the subject behaves in this experimental situation, then to discuss whether or not the experimental results differ significantly from his theoretical predictions, and finally to speculate on why the differences, if any, occur. Instead, the author's approach to theory construction is much more subtle. He begins with what seems to him to be the simplest conceivable model which could account for these elementary tapping tasks. Then, by a sort of dialectical process, he first elaborates the model slightly in a plausible way, so that it simulates more closely human performance, then complicates the experimental situation in order to test this new model, and so on. An example of the elaborations he introduces into the model is to assume that the memory of the time interval to be reproduced by the subject itself fades with time, whilst one of the experimental complications he uses is to vary systematically the inter-click intervals, so as to check the assumption that only the previous inter-click interval and the error in responding to it are used for predicting the next inter-tap interval. An additional task is devised in order to investigate whether, as would be expected, only particular parameters of the model are affected by divided attention. Throughout the development of the model, Michon uses both information processing and linear systems analysis, with the sequence of inter-click intervals being considered as input, and the sequence of inter-tap intervals as output. The model is finally presented as a computer program in order to demonstrate its internal consistency.

The one rather striking set of empirical results Michon has not yet incorporated into his model concerns the existence of temporal periodicities in the inter-tap interval distributions. To complicate matters, these periodicities vary with changes in the situation, but have a median frequency of about 12 cycles per second, a value which resembles some found in perception, such as the maximum subjective rate of perceived flicker. Perhaps this phenomenon is correlated to the findings that responses are made more frequently in one part of the motor cortex alpha cycle than during the rest of the cycle.

The author develops the mathematical techniques he uses from first principles, so that the monograph should prove comprehensible to most experimental psychologists.

T. SHALLICE

INDEX TO VOLUME 21

ARTICLES

Author	*Title*	*Page*
Anderson, J., Kearney G. E. and Everett, A. V.	An Evaluation of Rasch's Structural Model for Test Items	231
Audley, R. J. and Mercer, A.	The Relation between Decision Time and Relative Response Frequency in a Blue-Green Discrimination	183
Brelsford, Jr., J. W., Shriffrin, R. M. and Atkinson, R. C.	Multiple Reinforcement Effects in Short-Term Memory	1
Everitt, B. S.	Moments of the Statistics Kappa and Weighted Kappa	97
Falmagne, J. C.	Note on a Simple Fixed-Point Property of Binary Mixtures	131
Gregson, R. A. M.	Simulating Perceived Similarities between Taste Mixtures having Mutually Interacting Components	117
Jöreskog, K. G. and Lawley, D. N.	New Methods in Maximum Likelihood Factor Analysis	85
Kristof, W.	Structural Properties and Measurement Theory of Certain Sets admitting a Concatenation Operation	201
Massaro, D. W. and Moore, J. W.	Discriminative Binary Prediction with Reinforced Cue Identification	193
Maxwell, A. E. and Pilliner, A. E. G.	Deriving Coefficients of Reliability and Agreement for Ratings	105
Morton, J.	A Retest of the Response-Bias Explanation of the Word-Frequency Effect	21
Norman, M. F.	Slow Learning	141
Pask, G. and Lewis, B. N.	The Use of a Null-Point Method to Study the Acquisition of Simple and Complex Transformation Skills	61
Pike, A. R.	Latency and Relative Frequency of Response in Psychophysical Discrimination	161
Priest, H. F.	A Note on an Exception to the Rule of Decreasing Eigenvalues in Successive Factor Iterative Methods	265
Ramsay, J. O.	A Scoring System for Multiple-Choice Test Items	247
Ross, J. and Lumsden, J.	Attribute and Reliability	251
Shoemaker, D. M. and Osburn, H. G.	An Empirical Study of Generalizability Coefficients for Unmatched Data	239
Smith, P. T.	Cost, Discriminability and Response Bias	35

BOOK REVIEWS

Beth, E. W. and Piaget, J.	Mathematical Epistemology and Psychology	134
Cotton, J. W.	Elementary Statistical Theory for Behaviour Scientists	137
Hacking, I.	Logic of Statistical Inference	133
Magnusson, D.	Test Theory	136
Michon, J.	Time and Temporal Tracking	267
Runyon, R. and Haber, A.	Fundamentals of Behavioural Statistics	137

COMPUTER PROGRAM ABSTRACTS

Page, E. B., Fisher, G. A. and Fisher, M. A.	Project Essay Grade: A FORTRAN Program for Statistical Analysis of Prose	139

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Psychology (about 512 crown quarto pages per volume), *The British Journal of Medical Psychology* (about 350 crown quarto pages per volume), and the *British Journal of Social and Clinical Psychology* (about 320 crown quarto pages per volume) appear in quarterly parts, the subscription price of each, payable in advance, being £4 10*s*. per volume post-free ($16.50 in the U.S.A.). Subscriptions to any of these journals may be sent to any bookseller, or subscription agent, or to the Cambridge University Press, Bentley House, 200 Euston Road, London, N.W.1. American enquiries should be addressed to the Cambridge University Press, American Branch, 32 East 57th Street, New York, N.Y. 10022.

The British Journal of Mathematical and Statistical Psychology is issued twice a year, in May and November, in volumes of about 288 pages. The subscription rate, payable in advance, is £3 10*s*. per volume post free ($13 in the U.S.A.). Subscriptions should be sent to the Society's Agents, Taylor and Francis, Ltd., Red Lion Court, Fleet Street, London E.C.4; or to any bookseller.

The British Journal of Educational Psychology is issued three times a year, in February, June and November. The subscription is £2 a year post free, payable in advance, or £1 each number. The subscription price in the U.S.A. is $4.20. Subscriptions should be sent to the Subscription Department Manager, Associated Book Publishers, Ltd., Distribution Centre, North Way, Andover, Hants.

The series of *Monograph Supplements* issued in connexion with *The British Journal of Psychology* is now discontinued. Orders may be sent to any bookseller, or the Cambridge University Press, 200 Euston Road, London, N.W.1, from whom details of these publications may be obtained.

MANUSCRIPTS

for *The British Journal of Psychology* should be sent to
    Professor W. Sluckin, Department of Psychology, The University, Leicester;
for *The British Journal of Medical Psychology* should be sent to
    Dr J. Sandler, Room 101, Department of Psychiatry, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London S.E.5;
for *The British Journal of Social and Clinical Psychology* should be sent to
    Professor Thelma Veness (Social Psychology Editor), University of London Institute of Education, Malet Street, London, W.C.1, *or* to
    Dr R. R. Hetherington (Clinical Psychology Editor), 6 Abercromby Square, Liverpool 7;
for *The British Journal of Mathematical and Statistical Psychology* should be sent to
    Professor R. J. Audley, Department of Psychology, University College London, Gower Street, London W.C.1;
for *The British Journal of Educational Psychology* should be sent to
    Professor J. D. Nisbet, Department of Education, University of Aberdeen, Old Aberdeen.

SEPARATES

Contributors receive on application fifty copies of their papers free. Additional copies of separates may be obtained in units of fifty copies.

SUBSCRIPTIONS TO THE BRITISH PSYCHOLOGICAL SOCIETY

The subscription for Fellows, Associates and Graduate Members is £8 10*s*. (£7 10*s*. if paid by Banker's Order). They are entitled to purchase any or all of the following Journals at the special rates of £1 10*s*. per Journal per annum: *The British Journal of Psychology*, *The British Journal of Medical Psychology*, *The British Journal of Social and Clinical Psychology*, *The British Journal of Mathematical and Statistical Psychology* and *The British Journal of Educational Psychology*.

Information on membership and activities may be obtained from the Secretary, The British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1.

V. A
21 MAR 1969

THE BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY

Volume 21, *Part* 2, 1968

CONTENTS

NORMAN, M. FRANK
Slow learning *page* 141

PIKE, A. R.
Latency and relative frequency of response in psychophysical discrimination 161

AUDLEY, R. J. and MERCER, ANNE
The relation between decision time and relative response frequency in a blue-green discrimination 183

MASSARO, DOMINIC W. and MOORE, JOHN W.
Discriminative binary prediction with reinforced cue identification 193

KRISTOF, WALTER
Structural properties and measurement theory of certain sets admitting a concatenation operation 201

ANDERSON, J., KEARNEY, G. E. and EVERETT, A. V.
An evaluation of Rasch's structural model for test items 231

SHOEMAKER DAVID M. and OSBURN, H. G.
An empirical study of generalizability coefficients for unmatched data 239

RAMSAY, J. O.
A scoring system for multiple-choice test items 247

ROSS, JOHN and LUMSDEN, JAMES
Attribute and reliability 251

NOTES AND CORRESPONDENCE
A note on an exception to the rule of decreasing eigenvalves in successive factor iterative methods 265

BOOK REVIEW 267

Printed in Great Britain by
Taylor & Francis Ltd., Red Lion Court, Fleet Street, London E.C.4